

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Using Clickstream Data to Analyze Online Purchase Intentions

Ricardo Filipe Fernandes e Costa Magalhães Teixeira



Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Dr. Vera Lúcia Miguéis Oliveira

July 21, 2015

Using Clickstream Data to Analyze Online Purchase Intentions

Ricardo Filipe Fernandes e Costa Magalhães Teixeira

Mestrado Integrado em Engenharia Informática e Computação

July 21, 2015

Abstract

Nowadays, traditional business techniques are almost deprecated due to the insurgence of the world of online virtual shopping, the so-called e-commerce. This new, in many ways, uncharted territory poses difficult challenges when it comes to apply marketing techniques especially traditional methods, as these are not effective when dealing with online customers. In this context, it is imperative that companies have a complete in-depth understanding of online behavior in order to succeed within this complex environment in which they compete.

The server Web logs of each customer are the main sources of potentially useful information for online stores. These logs contain details on how each customer visited the online store, moreover, it is possible to reconstruct the sequence of accessed pages, the so-called clickstream data. This data is fundamental in depicting each customer's behavior. Analyzing and exploring this behavior is key to improve customer relationship management.

The analysis of clickstream data allows for the understanding of customer intentions. One of the most studied measures regards customer conversion, that is, the percentage of customers that will actually perform a purchase during a specific online session. During this dissertation we investigate other relevant intentions, namely, customer purchasing engagement and real-time purchase likelihood. Actual data from a major European online grocery retail store will be used to support and evaluate different data mining models.

Keywords: Clickstream, Web Usage Mining, Data mining, CRM

Resumo

Hoje em dia as técnicas de negócio tradicionais estão ultrapassadas devido à emergência de novos modelos de negócio, nomeadamente no espaço *online* através da Internet. Este novo espaço de comércio eletrónico difere substancialmente das atividades tradicionais que têm por base espaços físicos. Assim, torna-se imperativo que as empresas adotem novas estratégias e sejam capazes de compreender as motivações que guiam os compradores *online*, caso pretendam ter sucesso no competitivo ecossistema virtual.

Os *logs* dos servidores são a principal fonte de informação, sobre os seus utilizadores, que as empresas dispõem. Estes ficheiros contêm detalhes sobre como cada cliente navegou pela loja eletrónica, mais ainda, através destes dados, é possível reconstruir a sequência exata das páginas a que cada um acedeu. Este tipo de dados, conhecidos como dados de *clickstream*, são fundamentais para conseguir compreender o comportamento dos utilizadores. Aliás, a análise e exploração desta informação são essenciais para melhorar a relação com os clientes.

A análise de dados *clickstream* permite, acima de tudo, a compreensão das intenções que motivam os utilizadores a realizar determinadas ações. A percentagem de conversão de utilizadores é uma das métricas mais conhecidas e que se relaciona diretamente com as intenções dos mesmos. Nesta dissertação são também explorados outro tipo de intenções, nomeadamente, fatores relacionados com os utilizadores que passam a ser compradores e ainda com a probabilidade de compra em tempo real. São utilizados dados reais, provenientes de uma das maiores empresas europeias na área do retalho alimentar, para suportar e avaliar diferentes modelos de *data mining*.

Keywords: Clickstream, Web Usage Mining, Data mining, CRM

Acknowledgements

Prima facie, I am grateful to my family, namely to my parents and my brother, for all the support that was, is and always will be the main reason of my successes.

To my mentor, Dr. Vera Oliveira, I am thankful for her guidance and support, without which I could not deliver a dissertation with such quality. I would also like to thank all my teachers that, over the past five years, taught me according to the best standards, and made me the engineer I am today.

Before I enrolled this course, I was told that this academic experience would allow me to develop lifelong friendships. Five years later I can attest to that statement. I owe my friends all of the good memories and experiences that I lived throughout these years.

Finally, I wish to thank my institution, the University of Porto, namely, its Faculty of Engineering, for providing me the tools to learn and enjoy an academic experience.

Ricardo Filipe Fernandes e Costa Magalhães Teixeira

*“Chaos isn’t a pit.
Chaos is a ladder.”*

Petyr Baelish

Contents

1	Introduction	1
1.1	General Context	1
1.2	Problem Description	2
1.3	Contributions	3
1.4	Outline	4
2	Literature Review	5
2.1	Customer Relationship Management (CRM)	5
2.1.1	Classic CRM	5
2.1.2	Electronic CRM	6
2.2	Online Retail Market	8
2.3	Web Mining	9
2.3.1	Web Content Mining	9
2.3.2	Web Structure Mining	10
2.3.3	Web Usage Mining	11
2.4	Clickstream Data	11
2.4.1	Log File Structure	12
2.4.2	Preprocessing	15
2.4.3	Web Usage Mining: Previous Approaches	20
3	CLM Data Set	23
3.1	Introduction	23
3.2	Preprocessing	23
3.2.1	Preprocessing Pipeline	24
3.3	Metrics and Reports	27
3.3.1	Navigation Metrics	28
3.3.2	Trend and Traversal Reports	32
3.4	Discussion	37
4	Predicting Purchasing Engagement	39
4.1	Introduction	39
4.2	Problem Definition	40
4.3	Feature Engineering	40
4.3.1	Motivation	41
4.3.2	Data Set Restriction	41
4.3.3	Explanatory Analysis	41
4.4	Learning Models	48
4.4.1	Logistic Regression	48

CONTENTS

4.4.2	Random Forests	49
4.4.3	Evaluation Criteria	49
4.4.4	Results	50
4.5	Discussion	50
5	Predicting Purchasing Likelihood	53
5.1	Introduction	53
5.2	Problem Definition	54
5.3	Feature Engineering	54
5.3.1	Motivation	54
5.3.2	Data Set Restriction	55
5.3.3	Explanatory Analysis	55
5.4	Learning Models	60
5.4.1	Techniques and Evaluation Criteria	60
5.4.2	Results	60
5.5	Discussion	62
6	Conclusions	65
6.1	Clickstream Data Preprocessing	65
6.2	Clickstream Data Exploratory Analysis	66
6.3	Predicting Purchasing Engagement	67
6.4	Predicting Purchase Likelihood	68
6.5	Future Research	69
	References	71

List of Figures

2.1	Web Mining Taxonomy [CMS97]	10
2.2	Web Usage Mining Process [VMKR13]	12
2.3	Portion of a typical server log [Liu07]	13
2.4	Steps in data preparation for Web usage mining [Liu07]	16
2.5	Example of user identification using IP + Agent [Liu07]	18
2.6	Sample sessionization based on global time threshold of 30 minutes and local time threshold of 10 minutes [Liu07]	19
2.7	Sample for sessionization based on the navigation-oriented approach [Liu07]	20
2.8	Missing references due to caching [Liu07]	21
3.1	Histogram of the number of pages visited (left) and session visit time duration (right) for the CLM data set. All bounce visits were removed.	31
3.2	Histogram of the number of page requests per day of the week ¹ (left) and the number of page requests per day from users that logged and eventually purchased items (right).	33
3.3	Histogram for the number of page requests per hour (left) and the number of page requests per hour from users accessing the website through a mobile device ² (right).	34
3.4	Collection of box-plots specifying the in-session prevalence for each pageview, taking all CLM data set into account.	35
3.5	Percentage of sessions that requested a particular pageview regardless of the number of requests.	36
4.1	Histogram of the number of pageview requests for AIB and ANIB users.	43
4.2	Collection of plots assessing the quality of fitting a log-normal distribution over the empirical histogram, representing the number of pageview requests.	44
4.3	Simplified pageview session prevalence for AIB user sessions (top) and for ANIB user sessions (bottom).	45
4.4	Histogram of pageview prevalence for search related activities.	46
4.5	Bar plot specifying the pageview session span for both AIB and ANIB user sessions.	47
4.6	Box plots for the number of minutes spent per page for AIB and ANIB users.	48
4.7	Collection of plots for different performance metrics for predicting purchasing engagement. The top three plots refer to the evaluation of the logistic regression model, while the remaining bottom three relate to the evaluation of the random forest model.	51
5.1	Histogram of the probability of a basket addition at different session stages (number of pageview request).	57
5.2	Collection of plots assessing the quality of fitting a log-normal distribution over the empirical histogram, representing the moment of purchase.	58

LIST OF FIGURES

5.3	Collection of plots for different performance metrics for predicting purchase likelihood with an imbalanced data set. The top four plots refer to the evaluation of the logistic regression model, while the remaining bottom four relate to the evaluation of the random forest model.	61
5.4	Collection of plots for different performance metrics for predicting purchase likelihood with a balanced data set. The top four plots refer to the evaluation of the logistic regression model, while the remaining bottom four relate to the evaluation of the random forest model.	63

List of Tables

3.1	CLM Data Set Pageviews	26
3.2	CLM Traffic Metrics	28
3.3	CLM Metrics	30
3.4	CLM Conversion Metrics	32
3.5	CLM User Archetypes	33
3.6	CLM Landing Page Rank	35
4.1	The label, short description of the variables used in the model selection and the type of each variable.	42
5.1	The label, short description of the variables used in the model selection and the type of each variable.	56

LIST OF TABLES

Abbreviations/Acronyms

AIB	users that Added Item(s) to their Basket
ANIB	users that Added No Item(s) to their Basket
ASL	Average Session Length
AU	Anonymous Users
AUC	Area Under Curve
CEO	Chief Executive Officer
CLV	Customer Lifetime Value
CRM	Customer Relationship Management
e-CRM	electronic Customer Relationship Management
FMCG	Fast-Moving Consumer Goods
GDP	Gross Domestic Product
GIF	Graphics Interchangeable Format
GMT	Greenwich Mean Time
HTTP	Hypertext Transfer Protocol
IP	Internet Protocol
ISP	Internet Service Provider
JPG	Joint Photographic Experts Group
KPI	Key Performance Indicators
PNG	Portable Network Graphics
RM	Relationship Marketing
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
URI	Universal Resource Identifier
URL	Universal Resource Locator
W3C	World Wide Web Consortium
WWW	<i>World Wide Web</i>

Chapter 1

Introduction

This chapter presents the introduction to this research. It will briefly introduce its context, problem description and leading contributions. The last section is dedicated to explain how this thesis is organized.

1.1 General Context

In the last few decades, due to the booming of the *World Wide Web* (WWW) there has been a major change on how people interact with businesses. Nowadays, influenced by the social media and the fast lanes of communication provided by the Internet, traditional stores offer goods/services not only through the common physical channels, such as retail outlets, but also in online virtual stores, the so-called e-commerce. A recent study [NWW14] provided by an organization, Ecommerce Europe, that represents more than 25,000 European companies with a presence in the electronic commerce sector, reveals that the e-commerce business alone is responsible for 2.2% of the European GDP. Moreover, this study delves deeper into the online space and reports that there are 264 million (32%) e-shoppers in Europe, i.e. people that buy services/goods online, and these are accountable for 5.7% estimated share of online goods in total retail of goods.

Due to this increase in significance and market share, e-commerce companies have to adopt new strategies that fit the needs of the online customers [RF03]. This type of customers have different behaviors than the physical ones [CTH⁺10]. Furthermore, web users have a much easier job in comparing different companies through simple online queries, making it hard to maintain customer loyalty and retention the same way it is done in the traditional stores. Therefore, the online market poses a new challenge that requires new and bold Customer Relationship Management (CRM) strategies.

With the computer power flourishing in the early nineties, companies started to uncover the power of customer behavior mining. A new kind of directed marketing based on customer knowledge emerged in some of the world's biggest retailers. Tesco, for instance, is a major front runner

in applying data mining techniques to learn from what their clients buy [McE02]. Even when technology was not as accessible as today, Tesco knew that important features were to be extracted from what their customers were buying. They were pioneers in creating a loyalty card system that tracked all purchases made by users. On a six-year spree Tesco's Clubcard had amassed 104 billion rows of data. The main purpose of the card was to build a "segmentation and modelling system based on shopping behavior". The objective was met after a careful analysis of the collected data that culminated on a division into 25 categories that ranged all kinds of clients and allowed the retailer to practice target marketing strategies that appealed those in specific categories. It was reported that sales effectiveness of new stores increased by 50 percent since this method was applied [McE02].

On the customer side, this attention and reward for loyalty is welcomed, as this is nothing more than a modern version of what people used to enjoy back when the service provided by retailers was much more personal. Namely, when customers relied on a familiar salesperson to help them find just what they wanted [DDL11]. Nowadays, especially when it comes to the e-commerce, the sheer volume of data that is stored from the clients' interaction with the company can and should be used in order to simulate or, sometimes, outperform what used to be the familiar and personalized salesperson. In order to do so, e-commerce managers and marketers must depict a plan to improve the electronic Customer Relationship Management (e-CRM) which means delving into the field of data mining as well.

As the Internet works on a basis of interchangeable data, there are new data sources that companies ought to exploit. This data enables e-commerce managers to overview the business in ways that were not previously possible. Through an online store it is possible to track much more data that is a direct result of how the customer interacts with the company. The so-called clickstream data is key in understanding customer behavior and it is also the main source of information for the companies to adapt their service according to their audience [BS09]. In short, clickstream data, or clickstreams, is the common terminology for the collection of Web logs that compose the session of a specific customer on the company website. These sessions contain information regarding the path that a customer took through the website's structure, in other words, the sequence of clicks, hence clickstream, which a user performed and led him to a different point on the online store.

1.2 Problem Description

In order to build customer loyalty in e-commerce, companies must find an edge that will engage the consumer to repeat purchases in the future. The problem though, is that, unlike physical stores, customers are not bounded by the distance across stores. A mere few clicks can provide information from different sellers of the same product/service, therefore, companies ought to work twice as hard to strengthen the bond with their "e-customers" so that these do not defect.

One way to build this relationship is through the use of the so-called recommender systems [JZFF10]. This kind of technology ranges a wide variety of systems that can be more or less complex and intelligent. Recommender systems are a section of information filtering systems and

seek to predict the “rating” or “preference” that a user would give to a certain product or service. Companies have been using these systems to learn from their customers in order to provide a customized overall experience that helps them organizing the vast information spread across the website. The area of research for recommender systems is somewhat popular, mainly during the last decade, as there is several literature covering the topic. The state-of-the-art recommender systems are those that combine data from several angles, such as, demographical, clickstream and historical data [Les15].

Major retailers such as Amazon have developed algorithms that aim to develop offers and organize information according to each unique client, in a way that would be unfeasible to replicate in a physical store [LSY03]. However, with the recent developments in the fields of machine learning and data mining, new alternatives to analyze an exploit customer data ought to be discovered.

In this context, this thesis aims to develop a new methodology to support online CRM, with a focus on the e-grocery sector. The research applies data mining techniques to extract knowledge from databases containing clickstream data. Unlike the traditional recommender systems, the models developed in this study aim to use historical customer data to provide future insights that might help new marketing strategies increase customer retention. Moreover, there are two distinct problems that this thesis tackles. First off, there is the problem of discriminating users that are mere visitors from users that are potential buyers. Given that the industry conversion rate is bellow 6% [BLA⁺02], it is essential to identify this 6% potential audience and maximize their purchasing capacity in order to increase their customer lifetime value (CLV). A solid model that addresses this question and is able to classify each user as a possible buyer or non-buyer would be of great value to a marketing department. Furthermore, marketers could deploy targeted measures, early on the user session, and have a better chance of impacting the purchase decision. Additionally, there is a second problem that could have a deeper impact on how marketers respond to user behaviour. This question regards the likelihood of purchase on a specific page. In other words, given the domain of this study, online grocery retailing, for each request of each user, this hypothetical model predicts if any product is going to be added to the user’s basket. Putting it into perspective, the company behind the online store and their marketing team would have real time control over what and when their customers were or were not going to add something to their baskets. The possible marketing ramifications of this insight are not under the scope of this thesis, but, for instance, it would be possible to reorder product lists by profit, if the model yielded a high likelihood of purchase or, if that likelihood were to be low, a target pop-up window could be triggered to motivate that purchase.

Together, these two hypothesis outline the questions upon which this study will focus.

1.3 Contributions

In order to create a loyalty bond with online customers, businesses must to improve their websites to accommodate their needs, namely, help them organize and rank product information. This

is especially important when the underlying business is related to the food-retail market as the website aggregates information about thousands of different products.

During the past decade, several recommender systems have emerged, based either on collaborative or content-based filtering. The effectiveness of these systems has been proven along the years, with several companies, from different fields, succeeding due to them. Recommender systems and other data mining techniques can improve customer loyalty, increase cross-selling rates and increase the conversion rates of browsers into buyers.

Numerous literature has been published regarding these systems with main focus on customer clustering and product suggesting. This thesis goes further and aims to explore other forms of knowledge extraction. The research delves into online customer behavior and, using customer navigational data, i.e. clickstream data, tries to build a model that can output the likelihood of customer conversion while the customer is navigating the website. In other words, this research exploits the possibility of the clickstream data and the customer's intentions being linked.

In terms of academic contributions, this study is, to our knowledge the first to approach the online grocery retail market. On its own that is unique, specially because the grocery market, unlike other online markets, has intrinsic characteristics that need to be tackled differently. On the other hand, two models, regarding both hypothesis presented in the last section, were developed and each one of them is unique, both in terms of goals and predictor variables.

1.4 Outline

This thesis is concerned with providing further developments in the area of web usage analysis to explore web browsing behavior patterns. We will demonstrate our findings with a data set that was provided by a major European e-grocery. The structure of this thesis is as follows:

Chapter 2 provides a brief overview of the metrics/reports that can be obtained using clickstream data. We will introduce two types of metrics: for the website, and for a web session. These measures will be computed for the given data set and some results are reported through tables and graphs.

In Chapter 3 we focus on the development of two models that are able to predict customer purchasing engagement. We detail the process of extracting knowledge from data, also known as *feature engineering*, by using different graphical data representations. We also follow a specific evaluation process that helps quantifying the predictive power of both models.

Chapter 4 is structured very similar to Chapter 3 as we detail development of the same learning models but, this time, in order to predict real-time purchase likelihood.

Chapter 2

Literature Review

This chapter is dedicated to literature review. Background work will be reviewed, not only from computer science, but also from business and marketing. Firstly, the reader will have a deeper understanding on how current techniques are being applied to physical and electronic markets. Then, the field of Web Mining will be introduced and several known techniques detailed. The last part of this chapter is dedicated to review different approaches to related problems that were found in the literature.

2.1 Customer Relationship Management (CRM)

In order to successfully understand why customer's intentions prediction systems are important, one must analyze some business related fields. Ultimately, the end-users of these prediction mechanisms are people that are not necessarily related to computer science, furthermore, those who would benefit the most belong to marketing departments. Therefore, it is important to understand some of the concepts that serve as guidelines for thriving businesses.

2.1.1 Classic CRM

Back in the middle of the twentieth century, mass production techniques and mass marketing changed the competitive landscape by increasing product availability for customers. However, this process fundamentally changed the relation, sometimes personal, that business owners and employees maintained with their customers. Clients lost their uniqueness and started to become a mere "account number", the same way shopkeepers also lost track of their individual needs as the market became flooded with products and service providers [CP03]. Nowadays, every company is fighting for a competitive advantage that will guarantee their futures. Some of these companies are winning this challenge through the implementation of CRM principles allied with the power of technology, allowing them to improve customer retention and loyalty.

The concept of Relationship Marketing (RM), first introduced by Berry [Ber83], attempted to bring companies closer to their customers. RM is not focused on simple transactions but rather on customer retention and the establishment of more complex and long lasting relationships. CRM is based on the concept of RM and focuses on the underlying technology of customer management. Thus, CRM can be defined as the process of using information technology in implementing relationship marketing strategies, with particular emphasis on customer relationships [Pen13]. A successful implementation of a CRM strategy tends to improve common aspects of businesses regardless their specific area, namely: lower cost of customers' acquisition, better customer services, substantial increase in customer retention and loyalty, simplified marketing strategies and companies' productivity increase [BT04].

The analytical CRM, most relevant to this thesis, is the branch of CRM that is tightly related to technology and information systems. This branch can be categorized into four dimensions:

- **Customer identification:** This dimension can be split into customer segmentation and target customer analysis. Customer segmentation is related to the process of clustering customers according to their characteristics, while target customer analysis involves the specification of the most attractive segments to the company [KMS04];
- **Customer attraction:** After the identification process, companies allocate resources to attract the most valuable segments. Here, the marketing department has a crucial role as it is the best positioned company asset to perform this task;
- **Customer development:** The main focus of this dimension is to increase transaction intensity, transaction value and individual customer profitability. The notion of Customer Lifetime Value (CLV), i.e. the total expected income from a customer, is key at this stage as it should be one of the most important metrics. A higher CLV is linked to customers that are more loyal and are expected to repeat purchases thus resulting in a good prospect for the company's future;
- **Customer retention:** This is arguably the most important dimension of CRM. To achieve customer retention one must consider customer satisfaction as the most influential underlying factor [KMS04]. Customer satisfaction can be defined as the difference between the customer's expectations with the actual perceptions. Higher levels of customer satisfaction imply a higher rate of customer retention and, consequently, loyalty. To achieve this goal without falling into the mistake of mass marketing and generalization, companies' need to focus their attention to those segments that yield a higher CLV [AB02].

2.1.2 Electronic CRM

In a certain way the e-commerce helped companies shift the focus from the product to the consumer, a crucial step towards a successful CRM implementation. Either by necessity or innovation, businesses started adapting and morphing their old habits and applied CRM techniques to the Web space.

The normal online customer does not perceive the sheer amount of products the common Web store offers. Amazon, for instance, is a Website where customers have a range of books to choose from, so comprehensive, that it could not be replicated by a physical store. This notion of dimension may seem unimportant but, in the Web space, diversity leads to success. Amazon has proven that the average consumer is highly customizable, that is, consumers value diversity. The Amazon Long Tail phenomenon clearly proves that standardization is not the answer to consumer demand. In fact, it is estimated that 30-40% of Amazon's book sales are represented by titles that would not be normally found in brick-and-mortar stores [BHS06]. In his publication *Mass Customization* [Pin99], Joe Pine already argued that companies needed to shift from the old world of mass production to the new world where "variety and customization supplant standardized products". While e-commerce has not necessarily allowed businesses to develop more products, it has allowed them to provide consumers with more choices.

This level of customization has its consequences, it asks the consumer to digest an enormous amount of information on a computer, tablet or phone screen, without the convenience and organized environment of a physical store. This new scenario has changed some of the traditional aspects of CRM, thus a new terminology should be applied, i.e. e-CRM. Just like traditional CRM, "e-CRM concerns attracting and keeping economically valuable customers and repelling and eliminating economically invaluable ones" [RF03]. On the other hand, to address this information overload, e-CRM techniques involve applying mass customization principles not to the products but to their presentation in the online store [SKR01]. Furthermore, companies should develop and perfect the means of conveying information according to different users' profiles, as this is the only form of creating a competitive advantage other than reduced prices.

On a virtual dimension customers have the chance to compare prices across a vast array of different stores with minimal effort involved, therefore the churn rate is likely to be higher. One way to prevent this and achieve mass customization in e-commerce is the use of recommender systems. Recommender systems are used in e-commerce websites to suggest products to their customers and to provide consumers with information to help them decide which products to purchase. These recommendations can be an output as simple as a query based on the overall top sellers or a complex data mining algorithm that takes into account previous purchases and demographical data in order to predict future purchases. The recommendations themselves include suggesting products to the consumer, providing personalized product information and customization of queries to match profiles [JZFF10].

Recommender systems are comparable, to a certain extent, to marketing systems and supply-chain decision-support systems [SKR01]. Marketing systems support the marketer in making decisions about how the product should reach different audiences, usually grouped into different segments. By contrast, recommender systems can create a one-to-one, real-time evolving relation with every customer, theoretically allowing the number of segments to match the number of customers.

In terms of e-CRM these systems allow companies to establish a virtual connection with customers. Furthermore, depending on the complexity of the recommender, an e-commerce business

can have a more detailed insight of their audience than a regular business due to the possibility of continued activity tracking [CLA⁺03]. In sum, recommender systems enhance e-commerce in three ways [SKR01]:

- **Converting browsers into buyers:** The sheer volume of information a user has to digest can make it hard for a user to find the product they are looking for. Recommender systems can help consumers find products they intend to buy;
- **Increasing cross-sell:** Recommender systems can improve the cross-sell rate by suggesting additional products. A common example is to suggest products based on the customer's cart;
- **Building loyalty:** Customer loyalty is a cornerstone in CRM and it is no exception when it comes to e-commerce. One might be tempted to agree with the idea that it is hard to ask for consumer loyalty in the web space, as it is virtually effortless to look for better deals across various websites. The truth though, is that customers tend to be more loyal than expected [RS00]. Moreover, businesses themselves have the perception that a viable future relies on the company attracting loyal customers that repeat purchases over time. Recommender systems improve loyalty by creating a value-added between the site and the consumer. Sites, i.e. online businesses, invest in learning about their customers, use different data sources and recommender systems to present a customized interface and structured information according to different profiles. Consumers repay these sites by returning to the ones that best match their needs. This relation is mutually beneficial for when consumers return to the site, as they experience a more accurate degree of personalization, thus strengthening the bond between the online store and the client[SKR01].

2.2 Online Retail Market

This research focuses on the online retail market, more specifically, the online grocery business, also known as e-grocery. Therefore, this section will delve into this segment and explain some of its nuances that are distinguishable from a regular e-commerce business.

E-grocery is a market that has been steadily growing during the past decade and projections forecast even higher market share in a near future. In fact, the global online grocery business is expected to grow 25% over the next five years, according to a study conducted by SyndicatePlus [Syn14]. In the same report, this e-commerce product supplier for e-grocery businesses, drafted some interesting conclusions after interviewing 250 people from different countries in order to assess the actual consumer perception. One conclusion of this study is related to the fact that people tend to buy, by far, non-perishable goods. This can be related to the mistrust in the delivery system and it is one of the major challenges these businesses are faced with: assuring and maintaining high levels of quality control throughout the delivery pipeline. Companies are also aware that if quality is not assured, customers will not repeat purchases which is a serious problem given that the customer acquisition cost ranges from \$200 to \$700 [SS08].

While companies are tuning the handling of the physical products, further improvements can be achieved in the online domain. As discussed in the previous section, mass customization is key in any thriving e-commerce business and there is no exception for e-groceries. In effect, related literature has identified two phenomena of strategic change in the supermarket industry: “grocerification” and personalization. For one, “grocerification” relates to the opportunity that online businesses have to offer a wider variety of products to the end consumer. In practice, “grocerification” is about applying principles of grocery retailing to other products/services, using the internet as a cheap medium to acquire customers [LY04]. On the other hand, personalization, as one of most discussed subjects of digital business, holds a unique role in strengthening the relationships between online supermarkets and their customers. Through this online channel, supermarkets are able to provide personalized offerings to different customers. Target promotions and category management can be achieved with virtually no cost, enabling the “optimization of wallet-share spending for customers” [LY04].

In sum, e-groceries’ success depends on a joint effort that links the actual delivery system, with high quality standards and the frontend website that must be highly customizable and ready to exploit customer data.

"E-grocery success factors include cost-efficient delivery models, omni-channel marketing strategies, full-service customer support and rich digital product content availability. But true success requires deep consumer understanding. Retailers and Brands must find out what the online grocery shopper wants; when, how and why." [Syn14, Pieter van Herpen, CEO]

2.3 Web Mining

Web personalization is defined as any action that adapts the information or services provided by a website, e-grocery under this thesis domain, to the needs of a particular user or group of users, taking advantage of the knowledge extracted from users’ past actions, such as, navigational behavior. Web mining is the application of data mining methodologies, techniques and models to extract knowledge from Web data so that Web personalization can be achieved. On the other hand, data mining is the analysis of large data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner[CTS00]. Web mining is a broad field of study as it can be decomposed into sub-topics according to the different types of Web data available, i.e. Web usage data, also known as clickstream data, Web content data, and Web structure data [CMS99]. Figure 2.1 depicts a diagram about the taxonomy of the Web mining with its different branches and the primary source of data for each branch.

2.3.1 Web Content Mining

Web content mining is the process of extracting useful information from the contents of Web documents. The content data, in a site, is the collection of objects and relationships that is conveyed

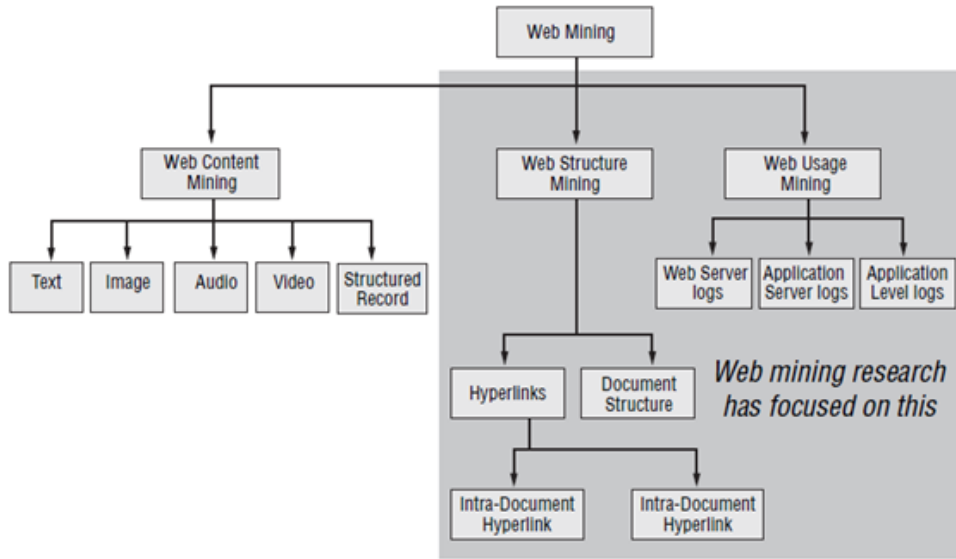


Figure 2.1: Web Mining Taxonomy [CMS97]

to the user. For the most part, this data is a combination of text, images, audio, video or even structured records such as lists and tables. Sentiment analysis [PL08], i.e. the application of text mining techniques to the web domain, is one of the most researched extensions of web content mining. With the emergence of different social networks, several literature has been developed under this topic. Twitter, for instance, has been a kind of laboratory for experiments for virtual political polls, through sentiment analysis of millions of “tweets” [TSSW10]; real time celebrity opinion [BMP11]; or simply general opinion mining [PP10]. While it is possible to extract information from other resources like images, videos and audio, the application of these techniques to web content mining has been limited.

2.3.2 Web Structure Mining

Web structure mining data refers to the information available in the inter-page linkage structure among web pages, as well as intra-page linkage structure within a page. The structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting linked pages. Structure mining is the process of discovering structure information from the web, and can be further divided according to the specific structure information used, namely, either through hyperlinks or document structure.

Hyperlinks are structural units that connect the World Wide Web, in short, they work as references for one Web location to another, either internal or external. On the other hand, there is research that focuses efforts into mining the topology of a Web page itself [Liu07].

2.3.3 Web Usage Mining

Web usage mining refers to the automatic discovery and analysis of patterns in clickstream and associated data collected or generated as a result of user interaction with Web resources, typically, a Web server. The goal is to capture, model and analyze the behavioral patterns and the interaction between users and a Website. The discovered patterns provide a useful insight on customer behavior as they usually represent resources that are frequently accessed by groups of users with common interests [Liu07].

The primary data sources in Web usage mining are the server log files, which include Web server access logs and application server logs. Each click made by the user while navigating the Internet, corresponds to an HTTP request to the Website's server, and it produces a new entry in the server entry log. The structure of the log is fairly common across different servers as they share common attributes, such as: hitting date; time of request; HTTP status; number of bytes sent; download time; user's IP address; path to requested resource; request's status; HTTP method; browser and operating system (agent); referring web resource; and, if available, client-side cookies [RSK13]. This type of files, universally known as log files, are the primary source of data for Web navigational behavior mining.

In some cases and for registered users, additional data may be on the server's operational database as a result of user's profile registration form. User data may include demographic information, user ratings on various products, past purchases and explicit or implicit representation of users' interests. E-commerce exploits this data along with clickstream to have a clearer picture of user's behavior.

Following the standard data mining process, [HKP06], Web usage mining can be divided into three sequential stages: data collection and preprocessing, pattern discovery and pattern analysis. During preprocessing, data is cleaned, partitioned, merged and transformed into data structures suited to be mined. Pattern discovery involves the application of data mining algorithms such as: statistical modelling, clustering and classification, association rule generation, sequential pattern generation and forecasting [Nga05]. Pattern analysis studies the potential predictive benefits of applying the mined patterns to the actual business environment. Figure 2.2 depicts a diagram that displays all three stages of Web usage mining.

2.4 Clickstream Data

Clickstream data or usage data collected automatically by the Web and application servers provides e-commerce businesses a unique chance to study their customer's navigational patterns and, by doing so, they will have a better insight on how to market their products. In order to become useful, clickstream data must be arranged in sessions that represent the path a particular client has made during his visit to the website. To exemplify what kind of information can be collected during a regular session, consider the typical behavior of a user who decides to purchase a product online. Firstly, the customer would reach the store's Website either directly, through the URL

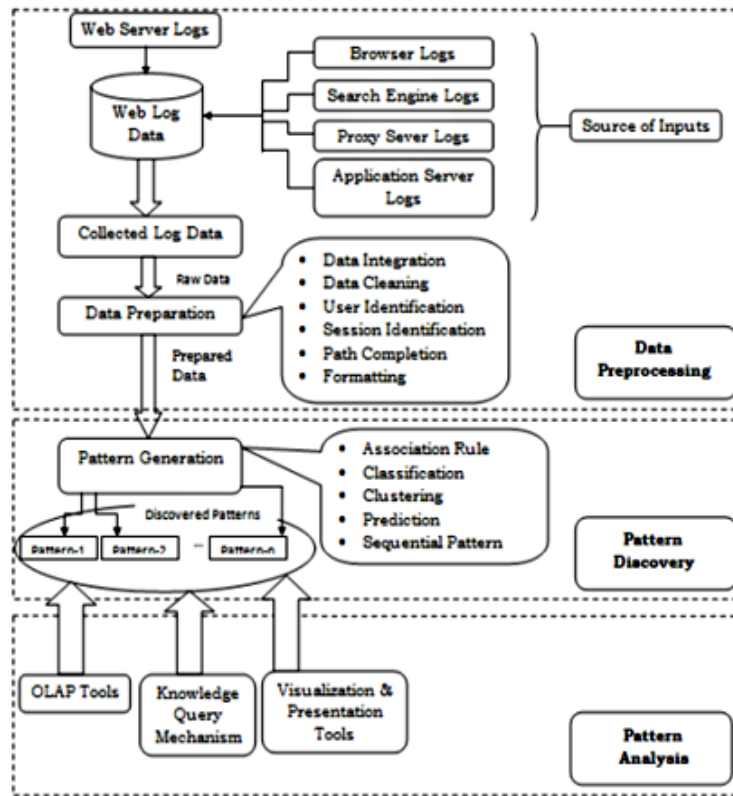


Figure 2.2: Web Usage Mining Process [VMKR13]

link, or indirectly, by searching for it on a search engine. Once there, he would start browsing by the product and click on different results until the right one was found. Then the user would add the item to a virtual cart and proceed to conclude the order by filling shipment and payment forms. Finally the process concludes once the client reviews and confirms the order, upon, usually, a confirmation email is sent. During this process clickstream data was recorded on the company's Web servers, allowing businesses to retrace and study customer behavior in order to improve marketing efforts [CTS00].

This section will delve deeper into clickstream data, namely, what are its attributes and what stages are involved in data preprocessing.

2.4.1 Log File Structure

Every action that any user performs on a Website is promptly recorded on server-side file, referred to as a Web log file or, simply, log file. Every single action matches a different line of the log, therefore a log may be perceived as a collection of different lines, corresponding to different actions, with several attributes. This file may be comma-delimited, space-delimited or tab-delimited. Although there are different norms for Web log files, the most relevant attributes are common among all. The World Wide Web Consortium (W3C) has published a standard, known as W3C logging

Literature Review

1	2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://dataminingresources.blogspot.com/
2	2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://maya.cs.depaul.edu/~classes/cs589/papers.html
3	2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200 318814 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1) http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey
4	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/
5	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html
6	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/header.gif - 200 6027 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html

Figure 2.3: Portion of a typical server log [Liu07]

that should be adopted by all parties. As follows, a brief description of the different fields which exist in a typical file is presented.

Figure 2.3 depicts a fragment extracted from a Web log where all the fields mentioned previously are present.

2.4.1.1 Remote Host Field

This field consists of the internet protocol (IP) address of the client that made the request. An IP address is a numerical label inherent to each device that uses the Internet Protocol for communication through the network. IP addresses serve two main purposes: host identification, indicates what the visitor seeks and location addressing, helps find out where it is [Dou92]. IP addresses are usually represented by dot-decimal notation, four numbers each ranging from 0 to 255. A typical IP address in the log entries would be *182.221.1.168*.

When it was conceived, the IPv4, the most common version of IP addresses, did not anticipate the sheer amount of devices that would end up joining the network. Therefore, nowadays there is a struggle to maintain the uniqueness of users through their IP addresses. Several Internet Service Providers (ISPs) have put in practice techniques where the pool of IP addresses allocated to them is shifting across different users, according to demand [BGL⁺09]. For Web usage mining this IP address management techniques can be hazardous since it is not possible to guarantee that the same device is accessing the Web server, just because they share the same IP address.

Nevertheless, for the experimental part of this research, IP addresses will be used as user identifiers as this is the only method one can do so with the given data set. Therefore historical purchases will not be taken into account, and consequently, user sessions do not need to be linked together.

2.4.1.2 User Name Field

The name of the authenticated user that accessed the server, i.e. this field only has any content for those users who have entered a password protected area of the Website. Although useful, especially for user's unique identification, this field is rarely used because the sensitive information is not properly encrypted. Therefore, anonymous users are normally indicated by a hyphen.

2.4.1.3 Date and Time Fields

The date and time of the local server for each request is recorded in the file in Greenwich Mean Time (GMT). These two fields can either appear as different attributes or coupled together onto a complete ISO 8601 date plus hours, minutes and seconds. This attribute is very important because it allows for the relative estimation of time that each user spends on each Web page.

2.4.1.4 HTTP Request Method Fields

The primary purpose of Web logs was to debug and keep track of every method related to the Hypertext Transfer Protocol (HTTP), an application-level protocol for distributed, collaborative, hypermedia information systems [FGM⁺99]. There are usually three web log attributes that can summarize the HTTP protocol: the request method, the Uniform Resource Identifier (URI) and the protocol version. The most common request method is GET, which represents a request to retrieve data identified by the URI. For instance, "GET /about.html" represents a request to the Web server to upload the page "about.html" to the respective user. As seen in the previous example, the URI contains the page or document name and the directory path that is being requested by the client browser. Sometimes, HTTP requests require extra information that is passed as a query. This would happen if a client performed a search through the Website's search engine, for example.

Some of the HTTP attributes are extremely relevant for Web usage mining because they allow for the tagging of resources that a user is requesting.

2.4.1.5 Referrer Field

The site that the user last visited and provided the link to the current request. This field is represented by that site's URL because the source can either be external or internal. The referrer field has long been used for marketing purposes since it can track where people came from. If the user entered the Website's URL explicitly in the correspondent browser's input field, there is no previous reference and this field is filled with a hyphen.

2.4.1.6 User Agent Field

The user agent indicates the user's browser, browser version and operating system. Most importantly, this field can contain information regarding bots or web crawlers. A crawler is a program that visits vast pages across the Web in order to keep data up-to-date, depending on its purpose.

The most common kind of crawlers are those associated with search engines, these kind of bots navigate Websites in order to create updated entries in their databases.

Even though it might be tampered, this field can also be used to assess if the user is human, thereby making it an easy filter to clean raw log data.

2.4.1.7 Status Code Field

The status code field is embodied by a three digit number that encodes the server response to a certain request. Codes of the form 2xx, for instance, indicate that the request from the client was received, understood and completed [FGM⁺99]. This field can also be used to perform further filtering, as invalid or server failures are not relevant for mining.

2.4.1.8 Transfer Volume Field

The transfer volume field indicates the size of the document, in bytes, sent by the server to the client. This information is used by network supervisors to monitor the load of the Web server.

2.4.2 Preprocessing

Web log files are not formatted to be mined and cannot be directly used if one hopes to create a solid model. Raw Web logs have to go through different stages in order to become mine-able. For instance, the original logs contain several records of user requests for Web scripts or images, i.e. elements from the Website that are not relevant to clickstream mining. This section is dedicated to the review of all the steps that are involved in preprocessing. Figure 2.4 presents a diagram that visually details the sequence of stages for data preprocessing.

2.4.2.1 Data Fusion and Cleaning

In large-scale Websites, it is usual for the server to be a virtual entity that is spread across different machines. In some cases these systems use redundant data in order to improve workload distribution and overall efficiency. Data fusion has an important role within these scenarios as it merges all logs from different sources into a single file, ready for further processing.

The next step of preprocessing is data cleaning. It is usually site specific and it mostly revolves around the filtering of unwanted references to objects that are not important in Web usage mining such as style files, scripts, graphics or sounds [EY10]. This cleaning may also entail the exclusion of some Web log attributes that are irrelevant to data mining analysis (e.g. HTTP version, number of bytes used, etc.). Moreover, cleaning can also involve the filtering of single-page visits, i.e. users that only sent one request to the server. These one-page visitors are irrelevant to this research as no behavior can be extracted from that information.

Data filtering for cleaning purposes can be accomplished through the analysis of the requested URI field, namely, its extension. For example, in the e-grocery domain, graphical resources can

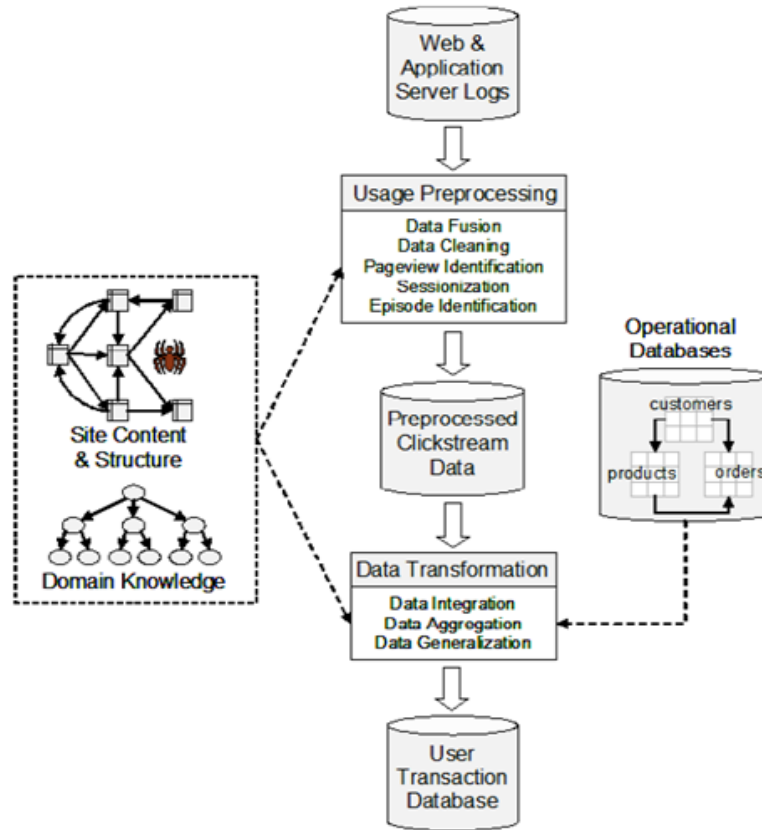


Figure 2.4: Steps in data preparation for Web usage mining [Liu07]

be considered irrelevant, therefore, requests for objects with extensions like JPG, GIF and PNG can be excluded from the log.

2.4.2.2 Despidering

Another important topic on data preparation is *despidering*, which entails the removal of references due to crawler navigation. With the everyday growth and constant update of the Web, search engines dispatch a series of automated software, the so-called crawlers or spiders, whose purpose is to search the Web space and provide links containing the information needs of customers [Cas05]. These bots' behaviors are quite distinct from humans, therefore it is crucial that records regarding their activities are removed from the data set, under the danger of tampering the learning process.

The most famous Web search engines' crawlers can be detected through the agent attribute on the Web log. Other heuristics can be applied to exclude these activities as "well-behaved" crawlers' first request is the "robot.txt" file under the domain's root. This practice is known as the standard robot exclusion protocol and it is designed to fasten the process of data cleaning. Even after these two methods of bot exclusion, the data set may still contain unwanted data, as an effect of bots that try to mimic human behavior. To detect and eliminate this data, one must apply more

complex methods. For instance an outlier statistical analysis of the data set can be effective as bots tend to access an abnormal amount of resources.

2.4.2.3 User Identification

The user's actual identity is not a requirement to perform Web usage mining. However, it is crucial to distinguish different users. When Websites have high traffic, Web logs can become unreadable to the human eye, because records from the same user would not be in sequential order. Moreover, one must be able to track the sequence of activities performed by the same user during different sessions, which is usually referred as user activity.

Ideally, user identification would be easily accomplished if the user provided login information, such as username and password, each time the Website was accessed. Unfortunately, the Internet's inherent design implies that most user online traffic requests are anonymous. In the absence of authentication mechanisms, the most widespread technique to distinguish different users entails the user of client-side cookies. Although effective, cookies are not completely reliable as users can disable them, due to privacy concerns.

Another method that solely relies on data already available in the Web log is based on IP addresses, as different IP addresses are likely to yield distinct users. Then again, due to the increasing number of internet users, ISP proxy servers are applying IP rotation methods to counteract the IP depletion phenomena [BGL⁺09]. For that reason, the same IP can belong to different users at different times. To overcome this problem, other heuristics can be applied, such as the pairing of both IP and agent fields to uniquely distinguish users [Jam11].

It is important to empathize that user identification is, in fact, machine identification, as the actual identity of the person who is using each device is unknown. It is perfectly possible to assume that the same computer is being used by several users, or even that the same user is accessing the Website through different devices (smartphone, laptop, desktop, etc.). The methods described are only heuristics that try to approach the data set with the most common use case, that is, each machine yields the same user.

Figure 2.5 presents an example of pre and post user identification using the IP address and agent fields as primary key.

2.4.2.4 Sessionization

Perhaps the most important step in Web usage data preprocessing is sessionization, i.e. the set of pages viewed by a particular user for a certain purpose. After user identification is completed, this stage performs a segmentation of user activity records from each identified user into sessions, each representing a complete visit to the Website. Without proper authentication systems, where sessionization can be performed by simply splitting the records according to each login/logout pair, one must rely on other heuristic methods. Different techniques have been studied by recent literature since the problem was first approached [CMS99]. New methods are still being published each one offering different nuances to their solution [CD08, DC11, KND13, RJR12, SS13].

Original Web log file					User Identification Output						
TIME	IP	URL	REF	Agent		TIME	IP	URL	REF	Agent	
00:01	131.2.13.94	A		Mozilla/4.01 (Win95, I)	USER 1	00:01	131.2.13.94	A		Mozilla/4.01 (Win95, I)	
00:09	131.2.13.94	B	A	Mozilla/4.01 (Win95, I)		00:09	131.2.13.94	B	A	Mozilla/4.01 (Win95, I)	
00:10	192.67.14.1	C		MSIE/6.10 (WinXP, I)		00:19	131.2.13.94	C	A	Mozilla/4.01 (Win95, I)	
00:12	192.67.14.1	B	C	MSIE/6.10 (WinXP, I)		00:25	131.2.13.94	E	C	Mozilla/4.01 (Win95, I)	
00:15	192.67.14.1	E	C	MSIE/6.10 (WinXP, I)		01:15	131.2.13.94	A		Mozilla/4.01 (Win95, I)	
00:19	131.2.13.94	C	A	Mozilla/4.01 (Win95, I)		01:26	131.2.13.94	F	C	Mozilla/4.01 (Win95, I)	
00:22	192.67.14.1	D	B	MSIE/6.10 (WinXP, I)		01:30	131.2.13.94	B	A	Mozilla/4.01 (Win95, I)	
00:22	131.2.13.94	A		MSIE/6.10 (WinXP, I)		01:36	131.2.13.94	D	B	Mozilla/4.01 (Win95, I)	
00:25	131.2.13.94	E	C	Mozilla/4.01 (Win95, I)		00:10	192.67.14.1	C		MSIE/6.10 (WinXP, I)	
00:25	131.2.13.94	C	A	MSIE/6.10 (WinXP, I)		00:12	192.67.14.1	B	C	MSIE/6.10 (WinXP, I)	
00:33	131.2.13.94	B	C	MSIE/6.10 (WinXP, I)	USER 2	00:15	192.67.14.1	E	C	MSIE/6.10 (WinXP, I)	
00:58	131.2.13.94	D	B	MSIE/6.10 (WinXP, I)		00:22	192.67.14.1	D	B	MSIE/6.10 (WinXP, I)	
01:10	131.2.13.94	E	D	MSIE/6.10 (WinXP, I)		00:22	131.2.13.94	A		MSIE/6.10 (WinXP, I)	
01:15	131.2.13.94	A		Mozilla/4.01 (Win95, I)		00:25	131.2.13.94	C	A	MSIE/6.10 (WinXP, I)	
01:16	131.2.13.94	C	A	MSIE/6.10 (WinXP, I)	USER 3	00:33	131.2.13.94	B	C	MSIE/6.10 (WinXP, I)	
01:17	131.2.13.94	F	C	MSIE/6.10 (WinXP, I)		00:58	131.2.13.94	D	B	MSIE/6.10 (WinXP, I)	
01:26	131.2.13.94	F	C	Mozilla/4.01 (Win95, I)		01:10	131.2.13.94	E	D	MSIE/6.10 (WinXP, I)	
01:30	131.2.13.94	B	A	Mozilla/4.01 (Win95, I)		01:16	131.2.13.94	C	A	MSIE/6.10 (WinXP, I)	
01:36	131.2.13.94	D	B	Mozilla/4.01 (Win95, I)		01:17	131.2.13.94	F	C	MSIE/6.10 (WinXP, I)	

Figure 2.5: Example of user identification using IP + Agent [Liu07]

Sessionization heuristics are categorized into two distinct sets: time-oriented and navigation-oriented. The time-oriented methods apply time-out rules to distinguish successive sessions. This heuristic has the underlying assumption that, if the same user has records across long periods of time, it is likely the he visited the site more than once. The standard time-oriented method defines a constant threshold, upon which if a user surpasses that time it is considered to end that session [CMS99]. More recent methods try to relax this hard constraint and produce a different sessionization using statistical quantities such as the average amount of time spent on each page [DC11]. Figure 2.6 portrays an example sessionization, obtained via two different time-oriented methods: global time-out, a threshold is defined for an entire session; local time-out, the threshold is defined for the total time spent between two subsequent requests.

Navigation-oriented heuristics use either the site structure or the implicit linkage structure captured in the referrer fields of the server logs. One method of doing so is to keep adding records to an existing session as long as the referrer field for that request was previously invoked earlier in the session. If not it is considered a new session. Note that with this heuristic it is possible that a certain request may belong to more than one open session, since it may have been accessed previously in multiple sessions. In this case, additional information can be used for disambiguation. For instance, the request could be added to the most recently opened session satisfying the above condition [Liu07]. Figure 2.7 illustrates an example of sessionization using a navigational-oriented heuristic. Comparing figures 2.6 and 2.7 it is clear that each method outputs different sessions. One must take the problem domain into account to choose the best fit.

2.4.2.5 Path Completion

Another fundamental preprocessing task, performed after sessionization is path completion. This step addresses the problem of client-side caching, which occurs when a user returns to a previously requested/downloaded page. The vast majority of Web browsers use caching techniques to save

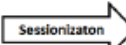
Original web log file				Time-Oriented						
				Global Timeout		Local Timeout				
TIME	IP	URL	REF		TIME	URL		TIME	URL	
00:01	182.22.3.18	A			1	00:01	A	1	00:01	A
00:09	182.22.3.18	B	A			00:09	B		00:09	B
00:19	182.22.3.18	C	A			00:19	C		00:19	C
00:25	182.22.3.18	E	C			00:25	E		00:25	E
01:15	182.22.3.18	A			2	01:15	A	2	01:15	A
01:26	182.22.3.18	F	C			01:26	F		01:26	F
01:30	182.22.3.18	B	A			01:30	B	3	01:30	B
01:36	182.22.3.18	D	B			01:36	D		01:36	D

Figure 2.6: Sample sessionization based on global time threshold of 30 minutes and local time threshold of 10 minutes [Liu07]

bandwidth and improve response times. On the other hand, whenever a user returns to the previous page, for instance, no request is sent to the Web server, therefore the server is unaware of these actions. However, missing references due to caching can be heuristically inferred through path completion which relies on knowledge of the site topology and referrer information from server logs [CMS99].

The most common case of missing reference occurs when the requested page is not directly linked to the last page. If so, one should look into the user's recent request history and assume that he backtracked to the closest page with a reference to the current request. As server logs only contain data about the time of the request, in cases of missing references it is also necessary to interpolate the time spent on each page. An approach is to assume that any visit to an already seen page makes it work as an auxiliary page, which is used to link the user to different pages. Knowing the Website's topology, the average reference length time for auxiliary pages can be used as an estimate of the access time for the missing pages [Jam11]. Figure 2.8 exemplifies a case where path completion was needed in order to have the correct clickstream sequence.

2.4.2.6 Data Integration

Web usage mining is even more effective when different data sources are combined and fed into data mining algorithms. For example, this research seeks to link clickstream data with customer conversion, i.e. try to extract a common pattern across those users who became buyers. These patterns are better derived if further information is available, namely, demographic user data, historical purchases or even product details. The data integration stage involves the combination of different types of data onto one entity with diversified and relevant attributes [KMPZ04].

2.4.2.7 Pageview/Transaction Identification

A pageview or transaction is a conceptual notion that is linked to some action performed on a Website. For example, clicking on a link, reading an article, zooming a picture, adding a product to the shopping cart, etc. The task of aggregating meaningful page references is called pageview

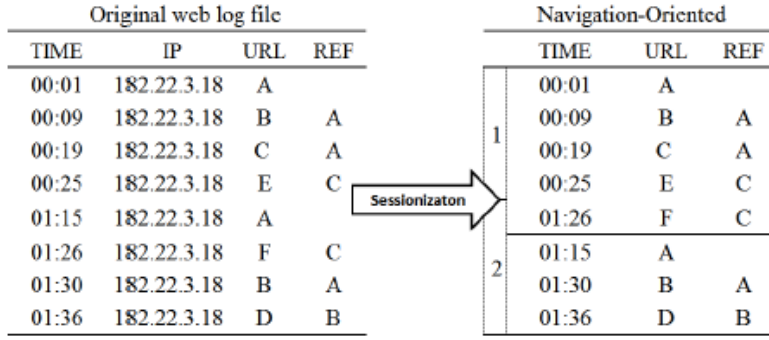


Figure 2.7: Sample for sessionization based on the navigation-oriented approach [Liu07]

or transaction identification [CMS99]. This identification is heavily dependent on the intra-page structure of the site, as well as on the page contents and the underlying site domain knowledge. For a static single frame site, each page request necessarily corresponds to a specific user action. However, nowadays most Websites have dynamic content and one transaction can correspond to a collection of requests.

It could also be relevant to create higher aggregation levels where seemingly distinct pages belong to the same transaction cluster because they fall in the same category. This process expedites the learning process as it decreases the overall entropy of the data set [Liu07].

This stage of preprocessing is strongly domain-dependent, however, its execution is crucial as one can only understand customers' intentions if there is knowledge regarding the objective behind each page request.

2.4.3 Web Usage Mining: Previous Approaches

As previously mentioned, understanding and predicting customer behavior and deploying marketing measures accordingly is of utmost importance to the success of e-commerce businesses [RS00, RF03]. The complex user behavior, mostly anonymous, combined with the high competition levels and low costs of switching product/service provider, accentuate the need to perform a thorough analysis of customer behavior [RS00]. Although still in its infancy, during the past decade, academic literature have been contributing with several publications in the field of Web usage mining. This section is dedicated to the review of some approaches that were published within this field of study and are more relevant to this dissertation.

In contrast with physical stores, online users are likelier to visit an e-commerce Website without any actual motivation due to the low effort involved. This is the main reason behind the extremely low conversion percentages [LT07], therefore, it is important to be able to distinguish the different possible customer profiles. [MF04] elaborated on this and developed a typology of Website visits, using navigational patterns, which identified four types of browsing strategies: directed buying, search/deliberation, hedonic search and knowledge building. Visitors following a directed

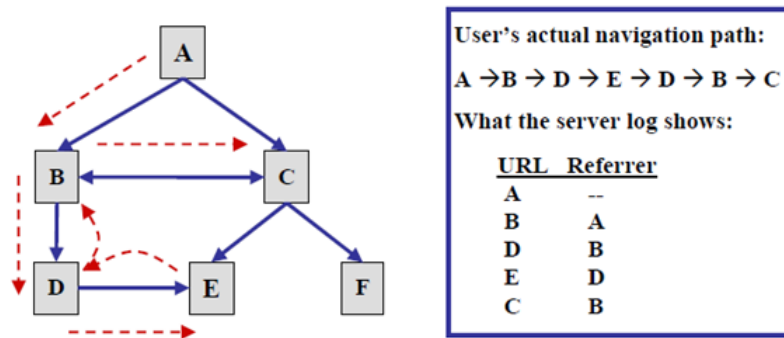


Figure 2.8: Missing references due to caching [Liu07]

buying navigational path intend to make a purchase and follow a focused and directed search pattern. Search/deliberation targets users that are also focused on their search, however, they have not yet decided which product to buy. Typical sessions belonging to these users are focused on one specific product category with many comparable similar products. Unlike search/deliberation, hedonic browsing visitors tend to portrait exploratory search patterns, they are more stimulus driven and have not yet decided in what product category to buy. Finally, knowledge building users show high levels of exploratory browsing and are not considering any concrete purchase [MF04].

Still under consumer behavior research, [MLSL04] tried to dichotomize customer behavior as either browsing or deliberation. This classification is similar to that of [MF04] as browsing is linked to an exploratory pattern with less purchase intention while deliberation is similar to focused search and is purchase oriented. The most relevant conclusion of their research is that visitors might alternate between different behavioral modes during a single session.

Due to this uncertainty, [MF04] developed a predictive model to accommodate for different types of behaviors. They were pioneers in feeding their model with clickstream data to predict purchase conversion with history of visits and purchases. During their research, the effect of visitors' return to the Website and their evolving visit behavior was studied. They concluded that more frequent visitors have a greater propensity to buy.

[VdPB05] studied the most relevant attributes for forecasting purchase behavior. They proposed a segmentation for the different kinds of variables: general visit-level clickstream behavior; detailed navigational path; customer demographics and historical purchasing behavior. The final model contained a set of variables that significantly improved performance, over previous studies. Moreover, they ultimately confirmed Moe and Fader's suspicions by showing the importance of within session detailed clickstream behavior.

One major drawback in comparing different methods is related to the tight correlation between the method's success and the problem's domain. For instance, [PC09] developed a purchase prediction model to the area of travel Websites. They justified customer conversion as a function of search motivation which also related to Moe and Fader's concept of browsing behavior. On the other hand, the study concluded that the number of pageviews tend to negatively impact customer conversion. This conclusion is refuted by [Ver12], whose empirical study has data from a retailer

selling consumer goods. Here, the number of pageviews is positively correlated with purchase conversion. These discrepancies are a clear indicator of domain bias, i.e. each domain entails specific characteristics. In this particular case, there is a clear distinction between an everyday consumer good and a travel appointment, in terms of price point and customer commitment.

All of the above-mentioned studies are mainly focused in understanding customer behavior and extracting relevant features upon which patterns can be drafted. However, there is another branch of literature that is tightly connected to computer science. These studies apply and derive different algorithms to that improve purchase conversion forecasting efficiency.

There has been a substantial amount of work on Web mining in computer science. The majority of this work is non-probabilistic and focuses on finding rules that describe common navigation patterns across users. [Nga05] and [NXC09] have developed a framework that reviewed and classified most of the relevant researches involving different algorithmic approaches. They performed a categorization of each study according to each of the four CRM dimensions, with the most relevant being customer retention. In terms of algorithms, there are two major branches of research in Web usage mining: clustering and classification.

Clustering techniques are usually performed using historical data and it is intended to segment customers into different groups. Each group shares some common features, e.g. the average time spent reviewing each product. [LC04] successfully developed a clustering model using association rules to mine customer knowledge over electronic catalogs that ended up being adopted by the company that was studied. Alternatively, [CK04] used a k-nearest neighbor algorithm to aggregate different clients upon which a recommendation system would deliver suggestions. Clustering methods are largely used to study customers, with the intention of deploying measures in the long term and to large groups. This is a common process for groceries as different segments are fed with different promotional coupons [McE02].

Classification methods are mostly applied to determine whether a visitor is a buyer or a browser. [KJSH06] used decision trees to classify and analyze customer value using a case study on a wireless telecommunication company. On the other hand, [CKLT03] have deployed support vector machines (SVM) to mine customer product rating in order to automatically generate target marketing offers. Perhaps the most relevant classification work for this research, due to its theme similarity, is that elaborated by [Ver12] who also used anonymous clickstream data to predict customer conversion. His work is very rich in terms of feature engineering and model approach. To date, his work is the only one that covers anonymous customer conversion. However, the specified model, lacks the ability to predict purchase likelihood throughout a user session. In other words, the model that this thesis explores aims to continuously track customer purchase likelihood throughout the session. Furthermore, his empirical study used data from an online clothing business while this research is targeted for the grocery retail sector. As previously mentioned, different scopes can have a tremendous impact in results, as the features are tailored for that specific domain. In this specific case, retail-based clickstream data is completely different from any other e-commerce business because sessions tend to be longer and users usually buy several items per session.

Chapter 3

CLM Data Set

In this chapter we aim to explore the metrics/reports that can be obtained using clicskstream data. We will develop this subject by measuring and analyzing a data set supplied by a major European retailer. This evaluation is the foundation for posterior models. It should be noted that we do not discuss all possible metrics, as that is beyond the scope of this study. Moreover, we will focus on a particular set of metrics that revealed to have an important impact upon the developed models. In fact, the web analyst does not necessarily measure all possible metrics, but only those that prove to help the company reach their goals. Further information regarding web analytics can be found in [\[DNB02\]](#) or [\[Cli12\]](#).

3.1 Introduction

Data is the single most important aspect of any data mining system [\[FPSS96\]](#). In order to extract knowledge from a data set, one must have an understanding of how it behaves and what are its Key Performance Indicators (KPI). This is arguably the most important step in every data mining pipeline. In fact, it is known that up to 90% of all the effort involved in developing a solid data mining model is spent on data analysis and preprocessing [\[TSK⁺06\]](#). During the development of this thesis we found this statement to be true, as we spent a large amount of time working over the data set both in terms of preprocessing and statistical analysis.

This chapter will approach and apply some known clickstream metrics to the available data set. But first, following up on the preprocessing pipeline introduced in section [2.4.2](#) we will describe how these methods were adapted and performed to set the data set for posterior modelling.

3.2 Preprocessing

Within this study we use data provided by a major European retailer. This clickstream data was retrieved from their commercial website that sells a vast collection of products that focus on the

grocery sector, but also allows consumers to purchase other goods such as clothes and electronic equipment. We will refer to this data as the CLM ¹ data set throughout this thesis. The CLM data set is a month's worth of user-session data from the company's servers. This data is both disaggregated and anonymous, in other words, we will not access registered user data such as demographic information and other session dependent attributes. The fact that we are working with anonymous data is important because the same model could be applied on other grocery retailers without having a market over fitting. On the other hand, if further user information was available, new features could be designed and the model would better fit the data.

The original data from the CLM data set lacked every step of preprocessing. This section's purpose is to briefly describe each step of the preprocessing pipeline described on section 2.4.2. We will re-enumerate each step and discuss its impact over the data set and its nuances and challenges.

3.2.1 Preprocessing Pipeline

As stated in chapter 2 on section 2.4.2, preprocessing the data is of utmost importance in order to build a solid clickstream prediction model. In truth, this prevalence of preprocessing is generalized to any classification model [LYL96]. The following enumeration presents the challenges and adaptations that we had to perform in order to preprocess the CLM data set compatible to the pipeline described on section 2.4.2:

- **Data Cleaning:** As previously stated, the CLM data set contained all requests sent to the server, the original log contains *212 675 331* rows of data. However, a vast part of these rows are linked to requests for Web elements that are not relevant for this study, such as, images, scripts and other media files.

After cleaning these unimportant elements the number of rows dropped to *33 105 835* rows, meaning that *84.43%* of the original data set was excluded at this first stage of preprocessing.

- **Despidering:** To perform despidering two techniques were applied. Firstly, the user agent field (see section 2.4.1.6) was used to filter some known Web crawlers such as the *GoogleBot* or the *BingBot*. On the other hand, there was a lot of traffic that being generated by other unidentified bots or high volume users and these were spoiling the data set. In order to identify and remove these entries, we had to perform a statistical analysis that considered the number of page requests per session. This analysis allowed us to target and exclude abnormally extensive sessions namely all sessions that performed over *300* page requests. We considered those as invalid user sessions. This threshold has a minor impact over the data set as it is set over 4 standard deviations away from the average, which corresponds to less than a *0.1%* cutoff.

¹CLM is a symbolic acronym for *Clickstream May*, because this data set relates to logs recorded during May of 2014.

- **User Identification:** The main goal of performing user identification is to analyze user recency, i.e. it allows the identification of repeated visits and derive other features such as haste². However, due to the nature of the website from which the CLM data set is from, that is, an online grocery retailer, where, we believe, people tend to purchase items on a monthly basis, we assume that this stage would not bring significant advantages. Furthermore, by assuming that each session is performed by a new user, we are extending the anonymity of the data set and thus contributing for a better chance of model generalization.

Therefore, user identification is performed on an intra-session level, with resource to the IP and user agent fields, but not on an inter-session level.

- **Sessionization:** Unlike standard server logs, the CLM data, due to confidentiality issues, was modified and some fields were either removed or masked. The most significant modification regards the referrer field (see Section 2.4.1.5) where there is no data. This absence of information impacts sessionization as it is not possible to know from which page users were coming from. Moreover, we had to assume that no client-side browser cache was used. Although this assumption is erroneous, its impact over the conclusions is not considered to be relevant as none of the model variables, described on Chapter 4, directly relies on its consequences.

Accordingly, sessionization was performed solely based on time constraints across requests. We applied a local timeout technique, instead of a global/local timeout technique. The reason behind this decision is related to the nature of the website itself, where, there are two major kinds of users, those who briefly visit the site and those who login and actually perform basket operations and, eventually, complete a purchase. These distinct behaviours yield contrasting log entries, as users who complete purchases tend to perform much more actions compared to users who do not. These divergent routines advise against global timeout strategies because sessions are very volatile. In consequence, we applied a local timeout technique when performing sessionization with a time threshold of 10 minutes. This threshold is standard in literature and there was no empirical study to determine it. This method amassed a total of 1 257 249 unique user sessions.

- **Path Completion:** Akin to sessionization, the path completion stage was affected by the lack of information regarding the referrer field. Given these conditions, as previously stated, we assumed that client-side browser caching was nonexistent. With no other methods or data to perform path completion, this stage was skipped.
- **Data Integration:** As we strive to implement a fully anonymous model, no data integration had to be done.
- **Pageview Identification:** Arguably the most time consuming preprocessing task regarded pageview or transaction identification. This task was also the one that introduced more

²The comparison of total time spent, compared to other sessions

CLM Data Set

Table 3.1: CLM Data Set Pageviews

Action	Meaning
Homepage	Requests for the website's homepage
Login	Request to authenticate a registered user
Logout	Request to terminate an authenticated session
Sub Category Menu	Request to display a descendant sub menu, e.g. clicked on <i>Bakery</i> to access a sub menu with different descendants such as <i>Bread</i> or <i>Cakes & Pies</i>
Category Page	Request for the display of all products under this category
Filter	Request to apply a filter to the present list of products
Search Results	Request for a text search (through the proper input box)
Autocomplete	Request for a list of terms that complete the current search term
Product Page	Request for the specific page of a particular product
Add to Basket	Request to add a specific product to the user's basket
Remove from Basket	Request to remove a specific product to the user's basket
Checkout	Request to checkout current basket and proceed to payment
Shipping & Payment	Collection of possible requests regarding shipping, scheduling and payment
Submit	Request for final submit of current purchase
Shopping List	Collection of possible requests regarding the creation and edition of personalized user shopping lists
User Profile	Collection of possible requests regarding different modifications to the user profile
Flyers	Request for a flyer with company promotions and discounts
Other	All requests that did not fit in the previous categories (0.68%)

knowledge about the specific data set domain onto the model. The primary assignment was to compile a list of all the accessed domains present in the data set, ordered by the number of requests to that page. Thus, the total 33 105 835 entries were grouped into 876 unique page requests.

With a determined set of requests, the next phase of this stage involved the understanding of company's website structure and what kinds of actions were possible to execute. With this in mind we went through most of the set of request and labeled each one according to its inherent action. This categorization was repeated until 99.32% of the requests were labeled. The remaining 0.68% were classified as undetermined/other requests and are mainly composed of unique urls to really specific pages within the site, rather than a broad user action. Through Table 3.1 it is observable that this analysis yielded 18 different types of transactions.

Although this procedure seems be company dependent, with chances of overfitting the future model, one has to keep in mind that these actions are common across all online grocery retailers. Therefore, although there is domain knowledge over this data set, we believe that this data set is both anonymous and generic enough so that we can draft conclusion over all

websites of this kind rather than this particular company.

- **Extra Preprocessing:** Apart from the standard preprocessing pipeline, there were other transformations that were applied to the CLM data set. These transformations are linked to some domain restrictions and were fundamental in order to keep data set coherence for further analysis. First off, we only consider sessions that request a minimum of five pageviews. This threshold relates to the business logic itself as the given company requires at least five requests in order to accomplish a full purchase. Since this thesis focuses on purchase likelihood, it is useless to analyze sessions presenting this minimal length.

The second main adjustment is also linked to the pageview identification stage, where, in order to keep data set coherence, we had to analyze the timings of each request and remove redundancies. For instance, the pageview identified as *Homepage* had various synonyms and, more importantly, was constantly observe as a sequence of simultaneous requests rather than a single one. If no actions were performed, further models would understand these sequences as actual user requests for the *Homepage*, when, in fact, they match a single user request but appear as a long sequence due to software design. In a sum, we had to analyze consecutive requests and understand if they match actual user requests or if they are a consequence of how the system was designed.

These extra transformations, the CLM data set was limited to a total of 25 691 403 rows that are grouped into 422 618 sessions.

3.3 Metrics and Reports

Web analytic reports from clickstream data using descriptive statistical methods can contribute to a better understanding of data, especially because it allows to investigate whether the website works towards the business objectives. These reports may include information such as the most frequent accessed pages, the average length of visit to a page, the average length of a specific path through a site, the most common entry and exit pages, the rate of visits with online purchase, etc. Even though these reports lack the depth of actually understanding specific customer behaviour, the resulting knowledge can be potentially useful for a better perception of the clickstream data set. Commercially, there are several companies that offer web analytics solutions and services available for clickstream data, including free vendors such as *Google Analytics* and *Yahoo Web Analytics*.

Nevertheless, there are some known metrics that help measure the attributes of web sessions. The variable most often used in the literature is the number of pages visited in a session [VdPB05]. If one has access to a clickstream data set that spans across a reasonable amount of time, the recency³, is an interesting metric that helps to understand customers' shopping habits [MF04, VdPB05]. The average time someone spends during a session on web pages and the total time spent at the site during the entire period of observation are among other time-related measures.

³Amount of time elapsed since the last visit.

Table 3.2: CLM Traffic Metrics

Variable	Frequency
No. of Visits / Unique visitors	422 618
Total no. of pages visited	25 691 403

This section is organized similarly to that displayed in the second chapter of Jamalzahed's work [Jam11]. Therefore, the metrics will be divided into two main sections: Navigation Metrics and Trend/Traversal Reports.

3.3.1 Navigation Metrics

Clickstream data collected automatically by application servers is the primary source of data representing the navigational behaviour of visitors. Depending on the goals of the analysis, this data can be transformed and aggregated at different levels of abstraction to provide metrics to infer about user's behavior. In this section we review some fundamental metrics often used in the web usage context.

3.3.1.1 Website Traffic

Perhaps the most common indicator in the web usage context is the traffic of the website, i.e. the amount of data sent and received by visitors to a website [Pal02]. This measure is frequently used to assess the overall popularity of a website, although it can be applied to specific sections or pages within the website. The following types of information are used when determining web traffic:

- **Number of visits and unique visitors:** The number of sessions that the website is visited over a specific time period, known as *number of visits*, is usually used as a web traffic indicator. Another, perhaps more important, derivation of this metric is known as *unique visitors* and it refers to the number of sessions originated from distinct users or machines. This distinction is made with the use of *cookies* that are unique for each visitor. These statistics are usually charted over periods of time, e.g. on a weekly basis, and their purpose is to indicate whether a significant change from the natural variability of the metric occurs.
- **Total number of pages visited:** The total number of pages visited, usually equivalent to the total number of user requests/actions, is another measure to determine website traffic. Coupled with the average size, in bytes, of a single page, this indicator can be used to estimate the load of a web service. Additionally, marketers can investigate the total requests for a specific page to measure the success of a new ad placement, for instance.

3.3.1.2 Website Stickiness/Slipperiness

According to the philosophy of CRM foundations, discussed on Section 2.1, more importantly than engaging website traffic is to encourage users to spend some time on the website and keep

them interested in it. This concept is usually referred to as *stickiness* in the web analysis context, or sometimes known as level of *engagement*.

Stickiness is frequently related to the profit of the website, as the likelihood of completing a purchase increases with the time spent on the website, assuming there that the website is optimally design to convey information. However, it should be noted that high values of stickiness is not always an advantage to the website. When users are navigating through *content pages*, which contain product or service information and descriptions, it is desirable for them to spend more time. Oppositely, for a *profile page*, such as registration page or shipping information page, the more time/clicks users perform, may indicate that they are not being efficiently guided through the process. There are multiple aways of evaluating stickiness, the following is just a short summary of some variables used within this study:

- **Bounce Rate:** An important measure that falls into this category of stickiness assessment is called *bounce rate*. The *bounce rate* refers to the percentage of visitors who come to the website but do not engage and leave the website after a few seconds or only visit a single page. Similarly to other measures, the *bounce rate* can also be calculated to particular sections or single pages within the web site, this can also be denominated as *page bounce*.
- **NPV:** The Number of Pages Visited (NPV) is a web site session is a popular measure to indicate depth of visit and stickiness. This metric can be generalized for the entire data set by computing the average number of pages visited (ANPV).

$$ANPV = \frac{\text{total number of pages visited}}{\text{total number of sessions}} \quad (3.1)$$

It is also helpful to analyze the histogram that represents the distribution of the NPV. This graphical representation of the data set allows for a further understanding of the underlying distribution that fits the NPV.

- **ASL:** Stickiness can also be calculated for a specified period using the Average Session Length (ASL) (see Section 3.2). This measure can be computed for the entire website, some sections or even specific web pages and, coupled with the NPV, is important to determine how fast users are assimilating information. Similarly to the NPV, the total time spent per session can be represented via an histogram that displays its real distribution.

$$ASL = \frac{\text{total time spent on the website}}{\text{total number of sessions}} \quad (3.2)$$

Table 3.3 displays the stickiness metric measured for the CLM data set. It should be noticed that, unlike the usual definition, we considered a *bounce visit*, all that did not requested over five pageviews. The industry value for a standard bounce rate definition, i.e. less than 5 seconds of activity, is 34%. The ASL value is higher than the overall online retail industry value, which is 9.5 minutes [Ret07]. However, this reference does not consider our restriction of 5 pageviews,

Table 3.3: CLM Metrics

Measure	Value
ASL	<i>10.58 minutes</i>
Bounce Rate	<i>66.0%</i>

at least, per session. Therefore, although useful, these values do not provide foundations for any conclusion.

Figure 3.1 (right) allows for a better comprehension of how much time users actually spend in each session. The ASL is the mean of this distribution but, although insightful, does not actually characterizes the data set in the best way due to the distribution's skewness. The median value, *5.85 minutes*, seems to represent a better estimate for a typical session, as it implies that half of the session last less then *5.85 minutes*. The same Figure 3.1 (left) depicts the distribution for the number of pages visited per session, where the average is *25.4 pages* and the median is *13 pages*. The correlation between these distributions is notorious as both have roughly the same shape and skewness.

3.3.1.3 Conversion

Conversion analysis tends to be the most business related metric as it directly relates to revenue and profit. A successful conversion in e-commerce occurs when a visit is guided by the website to purchase a product(s). The conversion rate, arguably the most used metric to measure conversion levels, is defined as the percentage of website visits that lead to an online purchase.

Several literature has been active on studying the underlying motives of purchasers in order to improve conversion rates [MF04]. If business managers understand the motives behind a purchase, they can deploy efficient strategies to convert more visitors into buyers.

It should be noted that conversion definition may be distinct for different types of websites or businesses. Concerning this study, a conversion occurs whenever a user adds an item to his virtual basket. Logically, one session may yield multiple conversions which should match the number of products in the basket.

There are numerous measures that evaluate visitor conversion. The following enumeration only contains a short set of indicators that are useful to estimate conversion levels for the CLM data set:

- **Conversion rate:** The most common way of measuring conversion is estimated by simply computing the percentage of sessions, for a specified time frame, containing user purchase.

$$\text{Conversion Rate} = \frac{\text{number of sessions that perform purchases}}{\text{total number of sessions}} \quad (3.3)$$

Typically, conversion rates fluctuate between *0.5%-8.0%* depending on the sector, target market or the definition of conversion. Within online retail financial services, for example, *1.0%-2.0%* would be typical with *2.0%* being very good [Kau07].

CLM Data Set

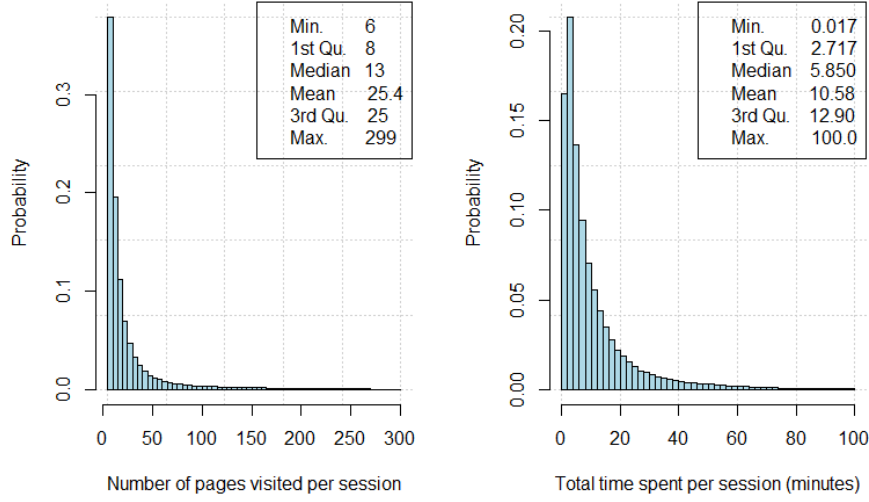


Figure 3.1: Histogram of the number of pages visited (left) and session visit time duration (right) for the CLM data set. All bounce visits were removed.

- **Number of pages to conversion:** The number of pages until conversion is an important measure because it portrays the efficiency of the website's purchase pipeline. On the other hand, comparing the number of pages to conversion of a particular session with the data set average (ANPC), can be used to assess hurry.

$$ANPC = \frac{\sum_{i=1}^n \text{number of pages to conversion for session}_i}{\text{total number of sessions with conversions}} \quad (3.4)$$

- **Average conversion time:** It is useful to provide a metric using the ratio of conversion time and session time, since this can be used to investigate at what point in a session a user decides to make an order. A desirable metric for an e-commerce website is given by the average amount of time shoppers spend on the website to buy an item online. A shorter conversion time shows a good performance by the website in terms of guiding visitors for e-shopping.
- **Time duration per conversion:** Staying on the website without performing any conversion is something that e-commerce websites try to minimize as it may result in slowing down the server. For an e-commerce website the time duration can be adjusted by the number of online shoppers. The Time Duration Per Conversion (TDPC) reveals the average amount of time elapsed per conversion.

$$TDPC = \frac{\text{total time duration}}{\text{total number of conversions}} \quad (3.5)$$

Table 3.4: CLM Conversion Metrics

Measure	Value
Number of Conversion visits (#)	45 801
Total Number of Conversions (#)	651 771
Average Number of Pages to Conversion (#)	15 15
Average Conversion Time (minutes)	6.51
Time Duration per Conversion (minutes)	6.86
Conversion Rate (%)	10.92

Notwithstanding, if the website is selling complex products/services, a longer visit may display that visitors are interested in obtaining as much information as possible. Usually this behaviour is linked to high priced items, such as laptops, where users tend to analyze product details more carefully.

Table 3.4 summarizes the set of conversion metrics applied to the CLM data set. The high conversion rate percentage is related to the definition of conversion. Since the data set refers to an online grocery retailer, it is expected that a regular buyer adds more than one item to the basket. Furthermore, the ratio between the number of conversion visits and the total number of conversions yields an average of 14.23 items added to the basket per conversion visit. The actual percentage of users that completed a purchase, regardless of how many items they bought, is 4.64% which is within industry standards.

Comparing the time duration per conversion and the average conversion time, one can conclude that the website is optimized towards conversion, as these values are relatively similar. Conversely, an average conversion time of 6.51 minutes reveals that the platform is not efficiently expediting users towards the products they need [NWW14].

3.3.2 Trend and Traversal Reports

Usually, metrics are only useful once it is possible to compare the data set from different points of view. Following that topic, this section will expand on the kinds of combinations that web analysts use in order to grasp their online audience.

3.3.2.1 Trends and Segmentation

Trend reports can provide the analyst with a better perception of any required metric as they show changes over a time period. These trends are usually segmented over different customer archetypes and provide useful insight over potential marketing strategies.

If the website has an authentication system, it is crucial to understand whether different kinds of customers behave differently. For an e-commerce online business a basic segmentation involves, for each metric, splitting the user base into buyers and non-buyers. Once these groups are properly identified, it becomes possible to study their differences and deploy congruent marketing campaigns with a better chance of success.

CLM Data Set

Table 3.5: CLM User Archetypes

Identifier	Meaning	Sessions (%)
AU	anonymous users	78.75
ANIB	users who did not add item(s) to their baskets	10.41
AIB	users who added item(s) to their baskets	10.84

Table 3.5 displays the different user archetypes within the CLM data set. During the month of activity from which the CLM data set was collected, a vast majority of sessions were anonymous (78.75%) and the remaining were almost bisected into ANIB users and AIB users.

The time-stamp available in the clickstream helps to find the visit-date, the date of a new web session in which the first request of the user's browser is sent to the server. Visit-date information enables us to derive other temporal session attributes such as whether the session takes place on weekdays, holidays, weekends, or any required period of interest. If the data set contains requests from different countries, by inferring coordinates from the IP field, one has to synchronize different timezones in order to accurately evaluate when requests were made.

One of the most important trend metrics relates to the measure of website traffic for different hours of the day or even different days of the week. This would show the peak time of the traffic and it can be exploited to find a suitable time to perform promotional campaigns or maintenance activities.

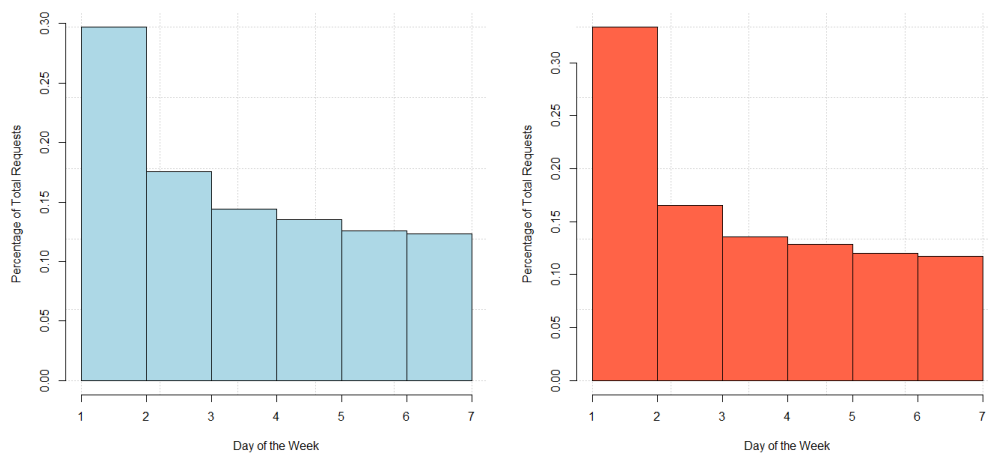


Figure 3.2: Histogram of the number of page requests per day of the week ⁴(left) and the number of page requests per day from users that logged and eventually purchased items (right).

There are several ways to segment the data and study distinct behaviors for each one, nonetheless, most of these segment analysis are inconclusive in terms of differentiation, as the shape of distributions tend to be very similar.

⁴Day one is equivalent to Sunday and the remaining are ordered naturally

⁵Including smartphones, tablets and smart TVs.

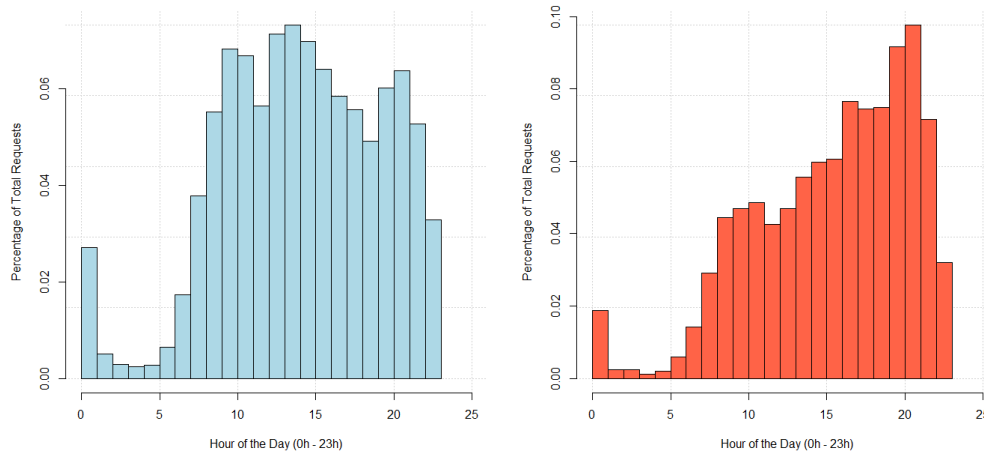


Figure 3.3: Histogram for the number of page requests per hour (left) and the number of page requests per hour from users accessing the website through a mobile device ⁵(right).

Figure 3.2 (left) displays an histogram of page requests grouped by days of the week. This distribution clearly highlights Sunday as the day with the most amount of page requests with the remaining days following an almost uniform distribution with a slight downward trend. Figure 3.2 (right) also represents the distribution of daily page requests but only focusing the segment of logged users that added item(s) to their baskets. Both plots yield similar distributions. Thus, a web analyst could conclude that there is no added value on targeting AIB users based on the day they are most active because, in essence, they follow the same pattern as the general user basis.

Figure 3.3 (left) shows how the number of page requests changes during the course of one day. This histogram reveals that page requests follow a similar distribution of that people tend to be awake. Figure 3.3 (right) depicts the same scenario but only considering the segment of the market represented by AIB users that access the website during Sunday. Although similar, it is noticeable that on the latter plot users are more likely to access the website during prime time while the general the page request distribution (left) shows no privileged time frame.

3.3.2.2 Traversal Reports

Apart from general information about a website and its sessions such as traffic, depth, time/date, time duration, depth of a session, etc., clickstream data also provides information about the sequence of web pages a user visits while browsing the website. In fact, after performing a pageview identification over the clickstream data (see Section 2.4.2.7), it is possible to analyze each session and understand how the user navigated through the website.

This insight uncovers a new type of analysis that analysts can execute. A common and simple diagnosis is to identify which are the most common landing pages, i.e. the first page that is requested within a session. This analysis could be useful to measure the success of a marketing campaign, for instance. If the company invested in a specific promotion or newsletter it is expected

that the page, where that promotion or newsletter is detailed, yields a higher rank on the landing page ranking.

Table 3.6: CLM Landing Page Rank

Rank	Landing Page	Requests (#)	Percentage
1	Homepage	314 305	74.37 %
2	Category Page	48 515	11.48 %
3	Product Page	17 272	4.09 %
4	Flyers	11 069	2.69 %
5	Search Results	6 862	1.62 %

Table 3.6 depicts the ranking of landing pages according to the pageview identification specified on Table 3.1. This rank reveals that the homepage is indisputably, the most requested first page and the top five landing pages are responsible for 94,25 % of all first page requests. A closer analysis of Table 3.6 may question how can visitors start their session through the *Search Results* pageview, as this action can only be accomplished once the website is loaded. However, this is a side effect of how we partitioned the session during the preprocessing stage. In fact, if a user loads the website's homepage and only request for search results 20 minutes later, this user session will be perceived as two separate sessions requesting one pageview each. For further details regarding *sessionization* see Section 2.4.2.4

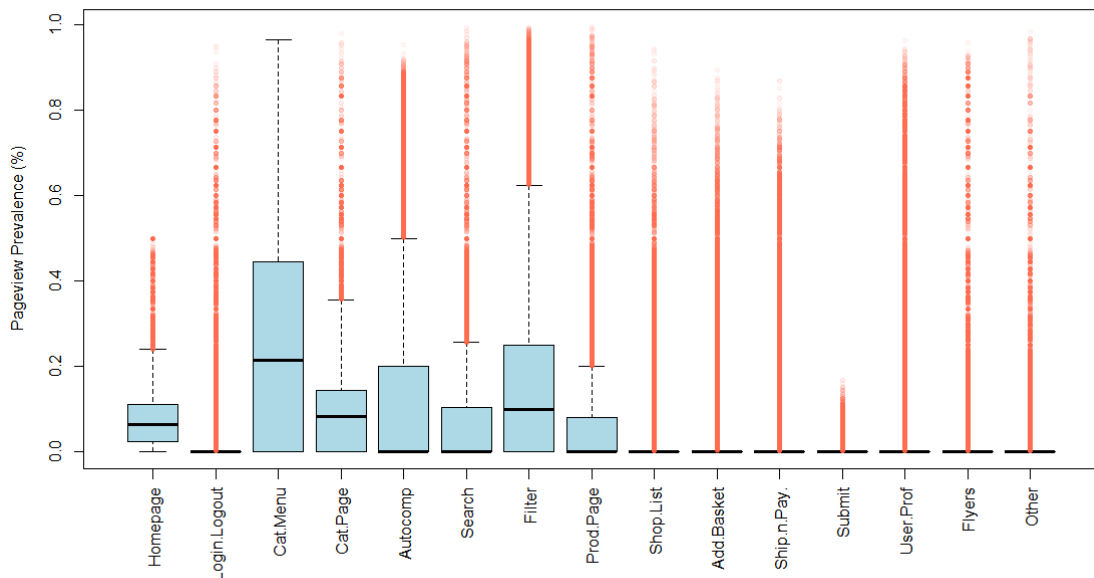


Figure 3.4: Collection of box-plots specifying the in-session prevalence for each pageview, taking all CLM data set into account.

Notwithstanding, there are other pageview analysis that allow for a better understanding of how the average session is divided into different user interactions. This sort of scrutiny is called

pageview session prevalence and it is fundamental for a real comprehension of user sessions. In essence, *pageview session prevalence* is the outcome of segmenting a user session pageviews set and compute how much influence of each pageview in the overall session.

Figure 3.4 depicts the pageview prevalence for the CLM data set. Since the vast majority of sessions are anonymous users (78.75%) it is natural that actions related to shopping tasks such as adding items to the cart or dealing with shipping/payment have a residual representation. Moreover, anonymous users are naturally linked to exploratory user sessions and logged users tend to display a more focused navigation pattern. This exploratory behavior translates to higher percentages of requests for content, hence the prevalence of pageviews such as *Category Menu*, *Product Page* or *Filter*.

Similarly to the *pageview session prevalence* analysis, where each user session is dissected into different pageviews, resulting on a broad picture of how each pageview impacts the session in terms of number of requests, it is also interesting to know the number of sessions that request each pageview, regardless of how many requests. We call this overview of the data *pageview session span* and its purpose is to highlight those pageviews whose business significance is relevant but are not requested that often. For instance, search related pageviews in any online store tend to be more requested than payment operations, yet the latter are more significant because they are directly related to the store's profit.

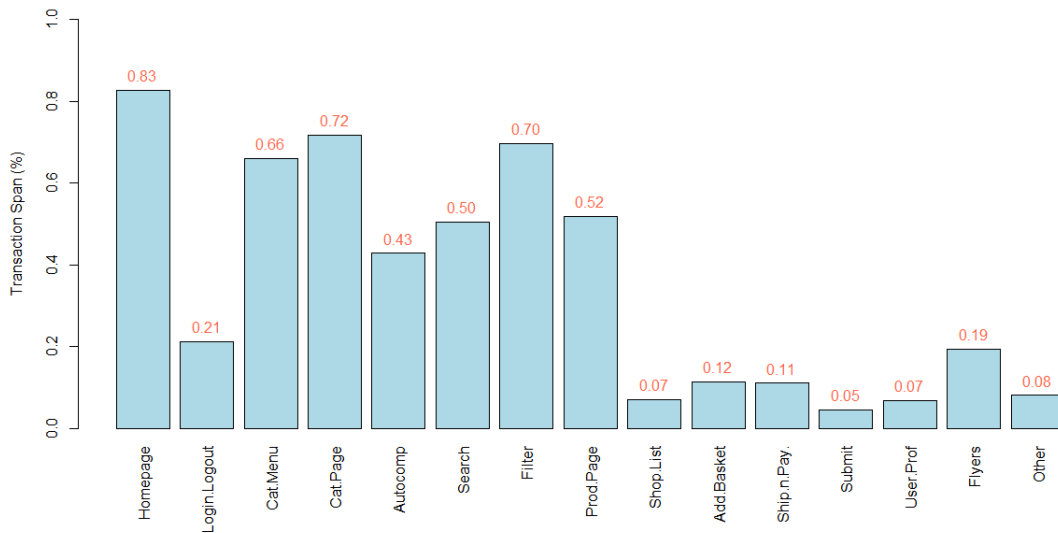


Figure 3.5: Percentage of sessions that requested a particular pageview regardless of the number of requests.

Figure 3.5 depicts the *pageview session span* for the CLM data set. Comparing this figure with the previous *pageview session prevalence* sufficient to understand its importance. In fact, we can identify some business important pageviews that have a residual session prevalence but,

a considerable amount of sessions are requesting them. The *Flyers* pageview, a page designed to convey advertising to end consumers, is a perfect illustration of this scenario as it has virtually no significance when analyzing session prevalence but, 19% of user sessions do request such pageview. In sum, both branches of pageview analysis are essential to actually understand how users explore the website.

3.4 Discussion

In this chapter, we detailed the process related with the preparation and measurement of the CLM clickstream data set. In addition to the common application of clickstream for website traffic engagement and depth, we used conversion information to report some KPI regarding the profitability. On the other hand, we introduced different kinds of segmentation strategies that allow analysts further means to identify specific user archetypes. Finally, we used the outcome of pageview identification to analyze and comprehend how a user session is split.

These measures will be used for different analytic and exploratory purposes throughout the thesis, namely during the *feature engineering* stage.

CLM Data Set

Chapter 4

Predicting Purchasing Engagement

In this chapter we introduce two models to determine the purchase engagement likelihood. A set of attributes derived from clickstream data are used as predictors. It should be noted that these models only use general anonymous navigational clickstream data. Whereas, in case of using registered customers' data one may access further information regarding each user (for instance, customer purchase history and customer demographics). This chapter's most relevant contribution relates to the initial exploratory data modelling and graphing because, to our knowledge, this is the first research that analyzes an online grocery retail data set with the purpose of purchasing engagement prediction.

4.1 Introduction

A well-known feature of online shopping is that visitors of e-commerce websites are rarely loyal to a specific website when searching for a certain product or service [VdPB05]. The search for the same product/service across multiple providers can be accomplished in a relatively short amount of time and, most importantly, with no costs associated. Thus, customers usually tend to look for the best offer they can find. Furthermore, this leads to a more intensive competition among companies that share the same market. On the other hand, the number of people that replace physical shopping with e-commerce has been growing [NWW14]. Therefore, companies are eager to develop methods that may help with online customer retention. One method for retaining customers is to target those visitors that are not expected to engage purchasing, and deploy target marketing early in their session in order to increase the probability of engagement [HB12].

Clickstream data obtained from online virtual stores can be used as a source of information with regards to customer's buying behavior. Clickstream data analysis yields a vast array of session variables with the potential of providing insightful details that separates online buyers and non-buyers. This analysis may provide a better perception of online buying behaviour, as it helps to improve the conversion rate by examining the motives for purchases [BS09]. It should be noticed

that the meaning of *purchasing behavior* is subject to different interpretations according to the underlying online business. For the purpose of this research, as mentioned upon the introduction of the CLM data set (see Chapter 3), a *purchase* occurs whenever a visitor/user adds a new item to his virtual basket. Therefore, our assessment of purchasing engagement relates to the moment a visitor adds his first item to his virtual basket.

Previous literature dealt with this problem under the domain of conversion rate analysis [MF04, BS09, VdPB05]. In essence, conversion rate and purchasing engagement are synonyms for most of the reviewed literature, as most data sets relate to ordinary e-commerce websites. On the other hand, this thesis explores the domain of online grocery retailing where, on average, users buy several items at once [NWW14]. Therefore, it makes sense to differentiate two distinct behaviors: purchase engagement (first basket addition) and purchase likelihood (posterior basket additions). This chapter concerns the modelling and prediction of purchase engagement.

4.2 Problem Definition

The CLM data set can be segmented according to the different user types that originate each session. Moreover, we have identified three different user types: anonymous users, ANIB users and AIB users. The problem of predicting purchasing engagement is the same as distinguishing between ANIB users and AIB users. Unlike anonymous visitors, both these users perform an authentication somewhere along their session. On the other hand AIB users, as the acronym entails, add item(s) to their baskets and ANIB users do not. This problem discards every anonymous user sessions because, the website's design requires visitors to authenticate before they can add items to their basket.

In order to determine whether an authenticated user engages with purchasing, one needs to review the sequence of requested pageviews and look for a basket addition. In other words, the doubt regarding the engagement of an authenticated user only lasts until the first basket addition is requested. Notwithstanding, from a business point of view, there is an interest to confirm the future engagement as soon as possible. With this knowledge, marketers would be able to deploy measures to increase customer engagement, thus boosting revenue [HB12].

The problem addressed within this chapter regards the classification of authenticated user sessions as ANIB or AIB. This classification is done for each request that the user performs from the moment he was authenticated. Therefore, this binary classification will only use a subset of the CLM data set as all anonymous sessions (78.75%) are not relevant.

4.3 Feature Engineering

In this section we use graphical representations to show how general anonymous clickstream data can explain the browsing behavior in terms of purchasing engagement. Specifically, we explore the engagement in authenticated user sessions for the clickstream data set, by investigating session attributes segmented according user type (ANIB or AIB).

4.3.1 Motivation

There is an extensive collection of literature that approaches different data mining and machine learning algorithms and their implementations. Despite the many contributions to theoretical learning models, the success of the learning process, regardless of the method used, is ultimately related to the quality of its predictor variables [Dom12].

Feature engineering is a crucial step towards a solid learning model as it is the process of using domain knowledge to create features that make machine learning algorithms work better. In other words, it is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data. Although fundamental, feature engineering still lacks the formality and academic importance that is given to the theoretical learning models [Bro14]. Feature engineering is even more important when dealing with problems that were not previously approached by any literature or when there is a limited data set, either by quantity or quality.

For the scope of this research, this process revealed to be fundamental mainly because the CLM data set, at first sight, lacked the necessary information to develop upon complex problems, such as purchase engagement prediction. The remaining part of this section will cover the process of understanding the data and crafting features that yield a simple and focused description of the underlying problem.

4.3.2 Data Set Restriction

The problem of predicting purchasing engagement only refers to a subset of the original CLM data set. Moreover, only sessions that perform user authentication (login) are considered. Therefore, the tabular data that is fed to the learning models will only contain rows that match requests made after the authentication was performed.

Since engagement classification is the objective of this chapter, AIB user sessions were trimmed in order to exclude all request after the first basket addition. Thus, AIB user sessions only include requests performed from the beginning of the session until the first basket addition request.

Therefore, from the original CLM data set, only 89 806 user sessions are considered. Given all in-session restrictions, the considerable data set amasses to a total of 1 167 482 requests. From this pool of sessions 20% were set aside to build an untempered test set, used to evaluate the learning models. The remaining 80% were used during the exploratory analysis phase and training of the algorithms. This subdivision is stratified, such that the percentage of AIB and ANIB user sessions is approximately the same as that in the initial data set.

4.3.3 Explanatory Analysis

The process of feature engineering is closely related to explanatory analysis. Furthermore, this analysis is fundamental because it allows for a deeper knowledge of the data set and its domain. During this section we will introduce every feature that takes part in the learning models. We will only focus on the features that were used by the learning models, nevertheless, the actual

Table 4.1: The label, short description of the variables used in the model selection and the type of each variable.

Feature	Label	Type
logNR	logarithm of the number of requests (log-Minutes)	Numerical (continuous)
searchPerc	percentage of session pageviews dedicated to search activities	Numerical (continuous)
userProf	whether any pageview related to <i>User Profile</i> was requested	Categorical (binary)
flyers	whether any pageview related to <i>Flyers</i> was requested	Categorical (binary)
other	whether any pageview related to <i>Other</i> was requested	Categorical (binary)
avgTPP	average amount of time spent per pageview (Minutes)	Numerical (continuous)
MarkDisc	Markov pageview sequence discrimination value	Numerical (continuous)

explanatory analysis was more extensive as the process of feature selection requires a lot of trial and error.

Table 4.1 depicts the set of features that feed the studied techniques. These features were chosen after an extensive process of testing different hypothesis. For the remainder of this section we will briefly explain why these features add value to the classifier.

4.3.3.1 Number of Pageview Requests

A key factor of customer behaviour on a website is the number of actions (usually the number of pages visited) made by visitors during a session on a website. Depending on the visitors' purpose, the number of requests can vary. Therefore, our hypothesis is that AIB users' pageview request distribution varies from ANIB users'. In order to assess the validity of this hypothesis we analyze the distribution of the number of pageview requests for each user segment. It should be noticed that this feature is dynamic, i.e. it changes as the number of requests within a user session increases.

Figure 4.1 depicts a back-to-back histogram of the number of pageview requests for both user segments. It should be noted that the distribution regarding the ANIB users was clipped for sessions that requested more that forty pages, so that the last column contains the accumulated number of requests from the clipped section. From this plot it is plausible to assume the validity of our hypothesis. In fact, if one takes a closer look at last half of both distributions, we observe that sessions with more than twenty requests are more likely to belong to ANIB users. Thus, to a certain extent, it is possible to distinguish sessions from the number of pageview requests. We believe that this occurs because the lack of purchase intentions of ANIB user makes them more prone to engage with an exploratory behaviour. On the other hand, AIB users are driven by purchases, for the majority of the cases, they will try to minimize the number of requests before adding the first item to their basket.

Another important data mining guideline regards the normalization of parameters. That is, the learning process is more effective when model features are bounded by the same numerical values [Kan11]. Features that represent probabilities or even percentages are naturally bounded as

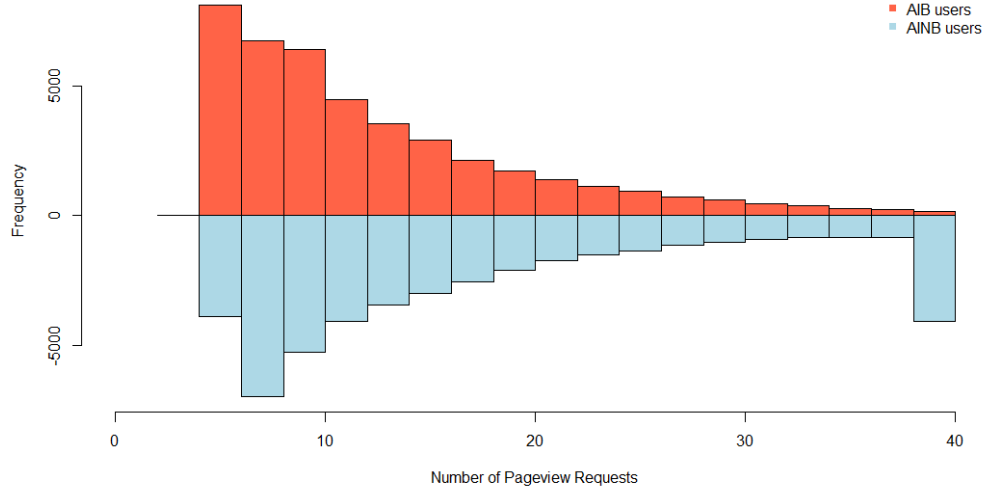


Figure 4.1: Histogram of the number of pageview requests for AIB and ANIB users.

they vary from zero to one. On the other hand, features such as the number of pageview requests are not only theoretically unbounded, as the number of pageview requests is controlled by the user, but also hyper-scaled when compared with percentage values. In order to restrain this feature into a controllable domain, we used the logarithmic value of the number of pageview requests as the input to the learning model. The logarithmic operation is effective because it is able to compress the given discrete distribution into a continuous one. The validity of applying this operator on this feature relates to the underlying mathematical distribution of the variable. Although distribution fitting is out of the scope of this thesis, Figure 4.2 depicts a collection of four plots that relates to the evaluation of fitting a log-normal distribution over the histogram of the number of pageviews for AIB user sessions. For further details on this issue see [RDTM79].

4.3.3.2 Focused Search

Another hypothesis that the exploratory analysis yielded relates to the fact that AIB and ANIB users seemed to navigate the website differently. For example, ANIB users seem more prone to perform user profile related actions, while AIB users are more focused on searching activities.

Figure 4.3 displays the *pageview session prevalence* box-plots (see Section 3.3.2.2) for both user segments. Instead of presenting a full granularity plot, in order to test our hypothesis, we merged all pageviews related with search activities (i.e. applying filters, navigating through menus, etc.) into one single action named *search*. From analyzing the plot we can conclude that AIB user sessions, on average, perform more search related activities than ANIB user sessions. Numerically speaking, 51.37% of all AIB user sessions requests, are related to search activities, while, for

Predicting Purchasing Engagement

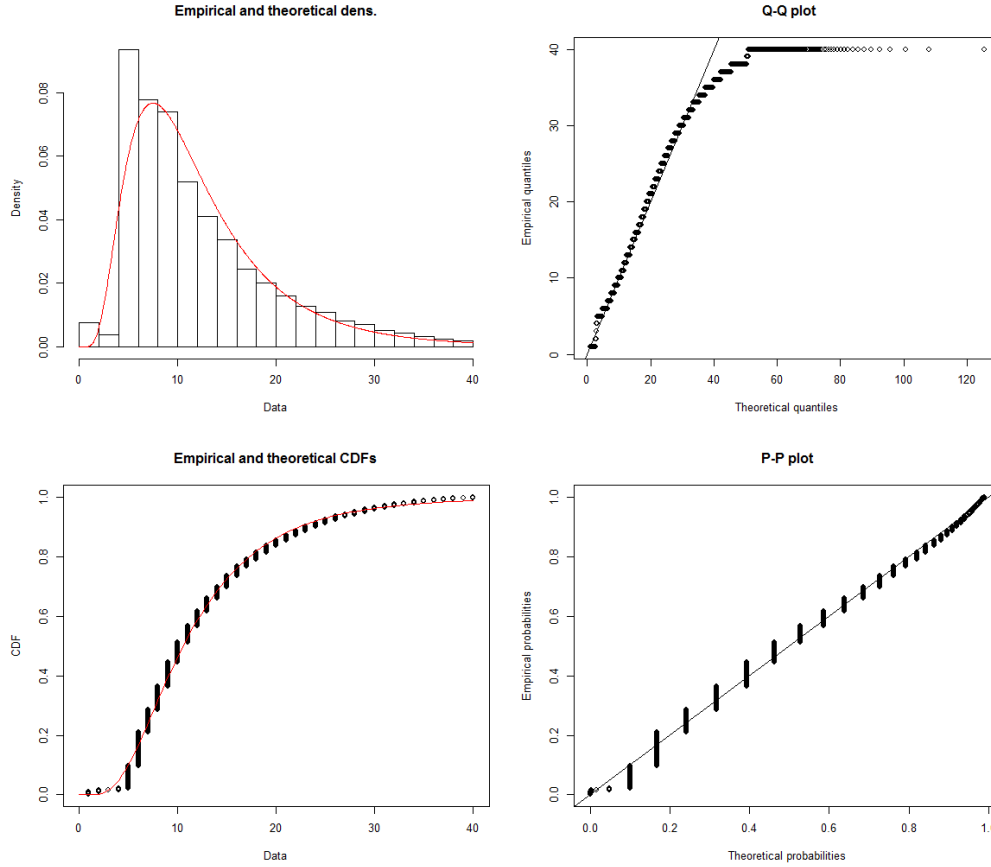


Figure 4.2: Collection of plots assessing the quality of fitting a log-normal distribution over the empirical histogram, representing the number of pageview requests.

ANIB user sessions, only 36.46% of the requests are focused towards the same end. Again, these statistics support the hypothesis that AIB users' sessions are more focused towards the same goal, they perform search actions to effectuate the purchase.

In order to further evaluate this statement, Figure 4.4 depicts the histogram of the search related activities for both user segments. This back-to-back plot reveals that both user segments have distinct underlying distributions for the same session prevalence attribute. This natural differentiation makes this feature useful for future binary classifiers. Similarly to other features, percentage of session requests related to search activities evolves during the course of the session.

4.3.3.3 Pageview Session Span

In order to develop relevant features, we have to delve into the motivations of different types of user sessions. For the AIB users, the process of authentication is normal because it is required in order to complete their purchase. On the other hand, it is not clear why ANIB users perform user authentication if they do not intend to purchase. One hypothesis is that ANIB users log-in, so that they can perform editions to their user profiles, such as, change of delivery address, elaboration

Predicting Purchasing Engagement

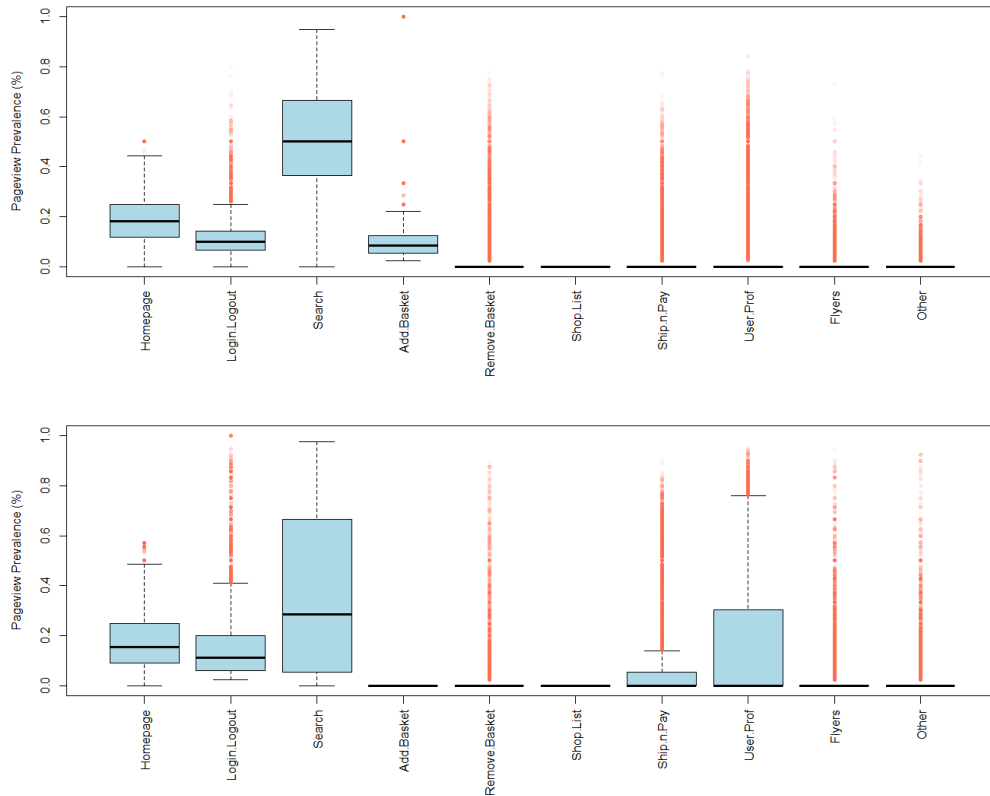


Figure 4.3: Simplified pageview session prevalence for AIB user sessions (top) and for ANIB user sessions (bottom).

of a shopping list, etc. As introduced on Section 3.3.2.2, in order to better understand the actions performed by each user segment we performed a *pageview session span* analysis for each user segment.

Figure 4.5 depicts the pageview session span for both user segments. Based on this plot we can confirm that AIB and ANIB users interact differently with the website. Moreover, our initial hypothesis is confirmed as ANIB users are prone to perform tasks that are not related to searching products. From the analysis of this pageview session span plot, we extracted three features based on the empirical difference between the two segments:

- **User Profile:** According to the information present in Figure 4.5 ANIB users are four times more likely to perform actions related to their user profiles. Moreover, 40% of all ANIB user sessions request these types of pageviews.
- **Flyers:** The same pageview session span revealed that 24% of ANIB users visit pages with promotional catalogs, also known as flyer pages, which is close to three times what AIB user sessions request.

Predicting Purchasing Engagement

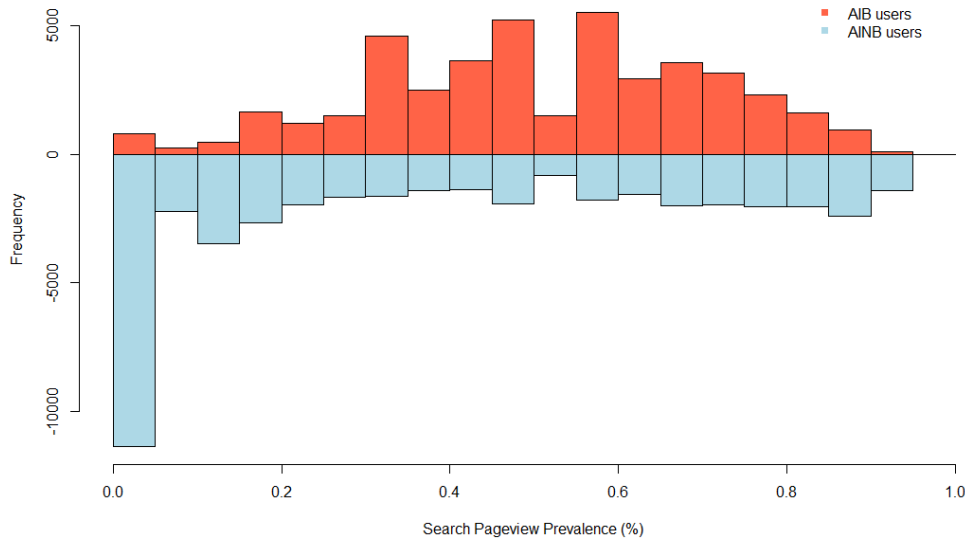


Figure 4.4: Histogram of pageview prevalence for search related activities.

- **Other:** We can also confer that 19% of ANIB users request pageviews that are not representative and fall into a category known as *Other*. Nonetheless, this percentage of requests represents almost five times more than the number of AIB user requests for the same pageview category.

These three features assume binary values, in other words, for a specific moment within a session each of the features will indicate whether that user has already performed any of those requests.

4.3.3.4 Average Time Per Page

Another clickstream data attribute is related to the average amount of time each user spends per page. Our initial hypothesis considered that AIB users were more focused towards purchasing items, thus spend less time per page.

Figure 4.6 depicts a box plot for each user segment, discriminating the amount of time spent per page. Although both distributions are, for the most part, overlapping each other it is noticeable that ANIB users tend to spend more time in each page than AIB users. For this reason, we considered the use of the average amount of time spent per page as a feature for the learning models. There is no need to scale this attribute as its natural distribution falls under a reasonable domain.

4.3.3.5 Pageview Sequence Likelihood

Thus far, every feature has been directly related to session metrics or specific actions. However, we assumed that the sequence of pageview requests can reveal whether a user is about to engage

Predicting Purchasing Engagement

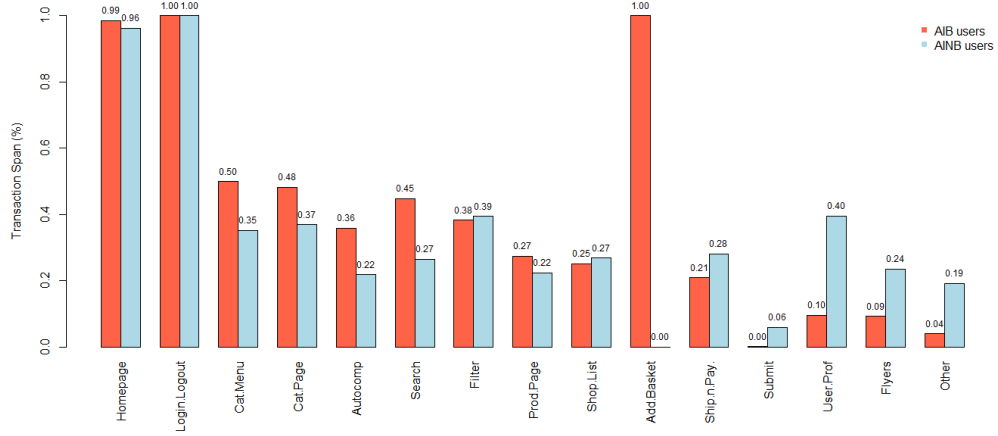


Figure 4.5: Bar plot specifying the pageview session span for both AIB and ANIB user sessions.

or not. In addition, we believe that AIB and ANIB users can be distinguished by their clickstream sequence. Therefore, we propose a measure of session similarity by using Markov for discrimination introduced by [DEKM98] and used, for instance, by [MVdPCeC12] for identifying churners and non-churners.

In a Markov model, the probability distribution of the next state depends only on the current state and not on the sequence of events that preceded it. Thus, each Markov process can be represented by means of a transition matrix. In the case of a process with N possible states, the corresponding transition matrix has a dimension of $N \times N$ states. Each element of the matrix, p_{ij} , represents the probability of the system evolving from a state i , in period t , to another state j , in period $t + 1$. In this research each state of the Markov process represents one pageview within a particular user session. In order to discriminate AIB and ANIB users using a Markov model, we assume that each user segment follow different Markov processes. Therefore, we build, for each population, a different transition matrix that reflects its specific pageview sequences. Following [DEKM98], we use these transition matrices to compute the log-odds ratio between the odds of observing a sequence x given it originates from the AIB users' population and the odds of observing sequence x given it belongs to the ANIB users' population:

$$S(x) = \log \frac{P(x | AIB \text{ user})}{P(x | ANIB \text{ user})} \quad (4.1)$$

The outcome of $S(x)$ allows the affinity of a visitor to be measured with respect to AIB and ANIB user sessions, by means of their specific pageview sequence. A positive ratio indicates that the visitor is not likely to engage with purchasing while a negative ratio means the opposite.

$$S(R_1 \rightarrow R_2 \rightarrow (\dots) \rightarrow R_n) = \sum_{k=1}^n \log \frac{P(R_k | AIB \text{ user})}{P(R_k | ANIB \text{ user})} \quad (4.2)$$

In order to compute the log-odds ratio for a sequence of pageviews, and not only a single

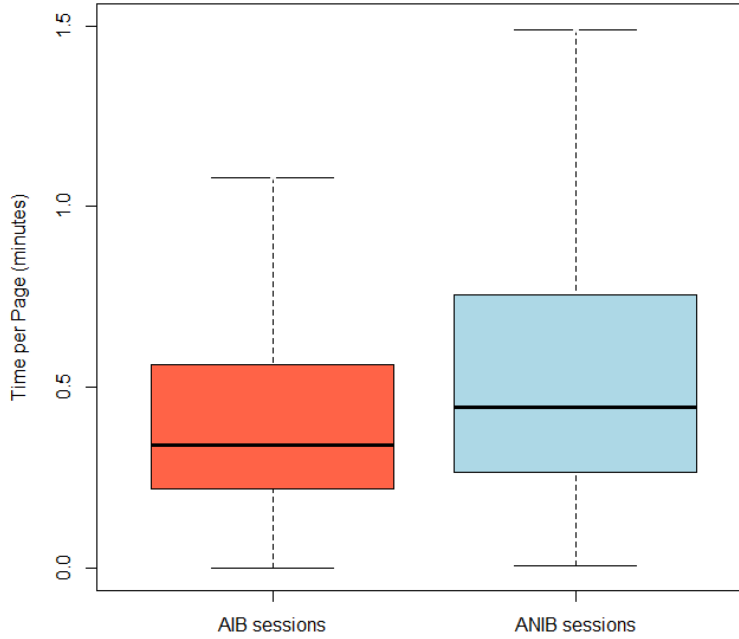


Figure 4.6: Box plots for the number of minutes spent per page for AIB and ANIB users.

request, we use the sum operation to join the log-odds of every transaction, as depicted on Equation 4.2. If this value is positive then that sequence of requests is more likely to come from a ANIB user while a negative ratio means the opposite.

In this study, we use the value of $S(x)$ as a feature for our models. We do not perform any domain scaling as the natural distribution of $S(x)$ is similar to the remaining features.

4.4 Learning Models

In order to test the predictive power of selected features, we submit our data set into two different learning algorithms: logistic regression and random forests. This section’s purpose is twofold, first we introduce each technique and review some related literature, then we specify the learning process, relevant parameters and evaluation metrics.

4.4.1 Logistic Regression

Logistic regression is a well-known technique, developed by [Cox58], which can be used for classification purposes. This is a form of regression which is commonly used when the dependent variable is binary, thus it is also known as binary logistic model. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by

estimating probabilities. The independent variables can be of any type [Agr96], but are usually continuous. The relationship between dependent and independent variables is not assumed to be linear. Instead, this technique assumes that the independent variables are linearly related to the logit of the dependent variables. Another upside of logistic regression is the fact that it does not require normally distributed variables.

This method has various applications spanning different fields such as medicine, politics or economics [BRD⁺00, HLC⁺84, CKT97]. It has become a standard classification method as it is easy to use and provides quick and robust results [VdPB05].

4.4.2 Random Forests

Random forests are an ensemble learning method for classification or regression, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (for classification) or mean prediction (for regression) of the individual trees. Unlike single decision trees, which are likely to suffer from high variance or high bias (depending on how they are tuned), random forests use several trees that vote upon which class is a better fit to a certain collection of features. This method was firstly introduced by [Bre01] as a method that combines the concept of *bagging predictors* [Bre96] and the random selection of features [Ho95].

Similarly to logistic regression, this method is widely used in academic literature. Random forests have very few parameters to tune and can be used quite efficiently with default parameter settings. This ease of use coupled with a strong predicting power, makes random forests a solid "off the shelf" choice.

This approach requires the definition of two parameters, the number of trees to be used and the number of variables to be randomly selected from the set of independent variables. For this study we will follow the recommendations reported by [Bre01] and consider a significant number of trees (i.e. 1 000 trees) and the truncated square root of the number of variables as the number of independent variables selected (i.e. 3 variables).

4.4.3 Evaluation Criteria

In order to measure the performance of each of the proposed prediction models, we compute the well known Receiver Operating Characteristic curve (ROC) and we analyze the Area Under Curve (AUC) as described in [HM82]. The AUC measure is based on comparisons between the observed class and the predicted class. The class is predicted by considering all threshold levels for the predicted values. An AUC close to 1.0 means that the model is a perfect classifier, while an AUC close to 0.5 suggests poor prediction capability as a random classifier is also expected to score 0.5 [HM82]. Moreover, we use the precision/recall curve in order to understand how this trade-off varies. In binary classification, *precision* is the fraction of retrieved instances that are relevant, while *recall* is the fraction of relevant instances that are retrieved. In simple terms, high precision means that the predictor returned substantially more relevant results than irrelevant, while high recall means that the predictor returned most of the relevant results. Precision and

recall are negatively correlated thus, the exact balance between these measures depends on the problem domain. Finally, we also use a simple yet informative accuracy curve as an evaluation metric. The accuracy is determined after the prediction models classify each case according to a certain threshold, i.e. 0.5 in this case. Then, accuracy is defined as the ratio between the number of correctly classified cases and the total number of cases to be classified.

4.4.4 Results

Due to memory constraints we had to limit the training data set to 50 000 entries out of the original 860 838 samples. In order to validate each model, we applied a validation strategy where we sampled 10 unique samples of 50 000 entries out of the whole training data set. Then, for each sample we fed our models with 80% of the data during the training phase and tested with the remaining 20%. This process of validation was carried out while we were developing and testing different features, so that, we avoided model overfitting.

Figure 4.7 presents the performance results of the two models using both logistic regression and random forests based on the predictions for the test set. Figure 4.7 depicts an agnostic evaluation of the prediction models in regards to the cut-off value as this is dependent on the marketers' intentions. For example, depending on the specific marketing campaign that would be deployed considering the model's predictions, one could value precision instead of accuracy.

From the analysis of Figure 4.7 we can conclude that anonymous purchase engagement prediction in this context is promising. The AUC values are high both when logistic regression and random forests are used. The results obtained also show that random forests outperforms logistic regression in all performance metrics. Moreover, the AUC value for random forests is close to 5% higher than the logistic regression's AUC.

Both prediction models were implemented using *R programming language* libraries.

4.5 Discussion

In this chapter we used the logistic regression and the random forests models to describe the association between general clickstream information concerning visits and whether a visitor will engage in online purchasing behaviour during his visit to the website. This model provides a new tool for an e-commerce web analyst that helps to infer visitors' behavior, and as a result, to improve online conversion rates.

The obtained results reveal that prediction models are effective in this domain. Unfortunately, the CLM data set does not contain the detailed clickstream variables and demographics that would increase the set of features and add more information to the prediction models. Nonetheless, the process of feature engineering revealed to be fruitful as, for instance, the random forest model is able to reach accuracy levels of 77%. On the other hand, these results show that anonymous clickstream data from online grocery retailers is enough to develop a solid model for purchase engagement prediction.

Predicting Purchasing Engagement

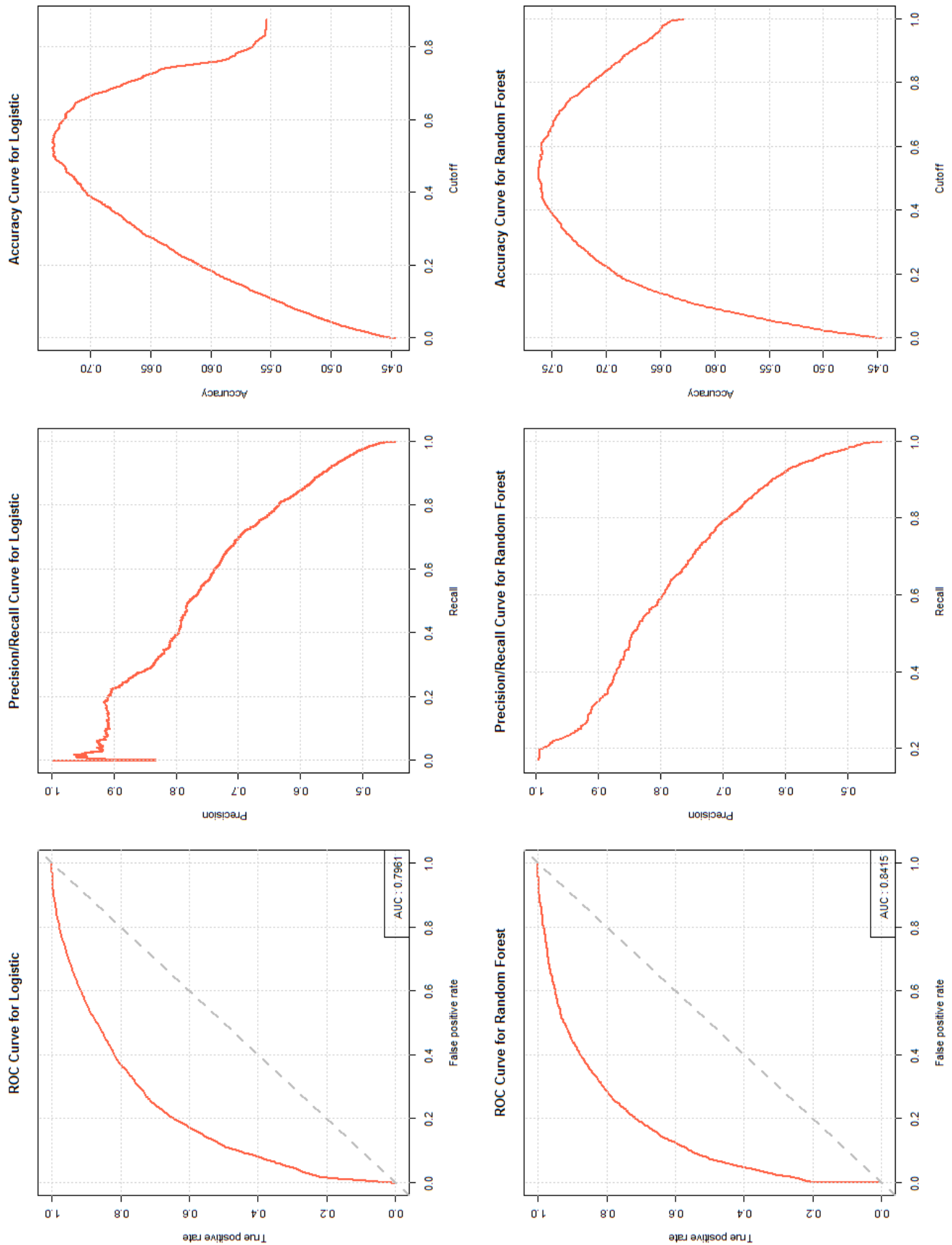


Figure 4.7: Collection of plots for different performance metrics for predicting purchasing engagement. The top three plots refer to the evaluation of the logistic regression model, while the remaining bottom three relate to the evaluation of the random forest model.

Predicting Purchasing Engagement

From the digital marketing point of view, these results are very encouraging as new methods of targeting customers could be derived from this solution. This research is also proof, at least for the company behind the CLM data set, that customers are predictable in terms of interactions with their website. In a web-focused marketing solution, this predictability might help to keep the customer in the site before leaving to find another competitor.

There are alternative approaches including neural networks, deep learning, categorical principal analysis and further decision tree-based methods which can be used to find the *best* approach for predicting customer engagement. Choosing the best approach, which is extensively discussed in the literature, would go beyond the scope of this thesis, and is not our aim. The logistic regression was chosen here for comparison reasons as most of our focus was dedicated to the random forests model. The predictor developed with random forests is solid and would be suitable for an extension of features if the CLM data set was also extended to include registered user information.

Chapter 5

Predicting Purchasing Likelihood

Similarly to Chapter 4, the present Chapter introduces two learning models to determine users' real-time purchasing likelihood. Following the outlined strategy, a series of features are derived from general anonymous in-session clickstream data. To our knowledge this is the first time that literature approaches real-time purchasing likelihood, i.e. being able to anticipate when clients will add items to their virtual baskets. Thus, the author considers the present chapter this thesis' most relevant contribution.

5.1 Introduction

On Chapter 4 it was introduced the problem of purchase engagement, i.e. given the e-grocery domain, predicting whether online visitors will add the first item to their basket and become an AIB user. Past literature approached purchase engagement prediction and purchase likelihood as two faces of the same coin. As previously mentioned (see Section 2.1.2), the nature of online grocery retailers differs from the most common e-commerce businesses. Moreover, the common e-commerce businesses focus on specific product segments such as clothing or electronics, i.e. not fast-moving consumer goods (FMCG). On the other hand, e-groceries' clients tend to mimic the same behaviour of a physical store and buy several items at once, which is contrary to casual e-commerce businesses. [Syn14] reports that, for instance, in England users tend to buy, on average, items from 6.6 distinct product categories within a single session. This distinction between e-groceries and other e-commerce businesses sets the foundation for the motivation behind the development of a model that predicts customer purchase likelihood.

Predicting purchase likelihood corresponds to the continuous prediction of basket additions. In other words, this model automatically follows user sessions, as they keep requesting new pageviews, and for each request it yields the likelihood of adding an item to the virtual basket, on the next request. Therefore, the most important outcome of this model is the power to dynamically anticipate clients' purchases.

In order to perform this analysis, we will continue to extract information from the CLM data set. Furthermore, some of the features we crafted during the development of the previous model (see Chapter 4) will be re-purposed to fit new requirements. Besides, we also re-analyze the available information in order to extract new and insightful features that complement our models.

5.2 Problem Definition

Assuming that the models developed on Chapter 4 are able to correctly forecast the engagement of a visitor, i.e. identifying them as AIB users, there is still room to explore user behavioural patterns, as AIB sessions, on average, last *30 minutes*, out of which only *35%* were considered to predict customer engagement. On the remaining *65%* of the session, AIB users assume a *purchase state*, where they search and eventually add different items to their baskets, just as they would do in a physical store. The hypothesis we explore in this chapter follows up on user engagement and seeks to understand if AIB users' actions are predictable to the point where we could know, given past pageview sequence, if the next request is going to add any product to the basket. Moreover, for each pageview request, performed by an AIB user, we predict whether the next request entails a basket addition. This means that, for every pageview request in moment t_i , given the sequence of pageviews already requested from t_0 up to t_i , these models predict the likelihood of adding an item to the basket on moment t_{i+1} .

Once again, web analysts would benefit from such prediction models as they would be able to control how each user would perceive content depending on their purchasing intentions. Moreover, if a marketer knew, before hand, that the user has the intention of adding an item to his basket after a certain request, then he would be able, for instance, to alter the ordering of the products, so that the most profitable ones would appear first.

5.3 Feature Engineering

In this section we review the CLM data set and extract attributes that are valuable for purchasing likelihood prediction. We use graphical representation to show how general anonymous click-stream data can explain the browsing behaviour leading to purchasing.

5.3.1 Motivation

The motivation and importance of feature engineering for purchase likelihood prediction are very similar to those that lead to the explanatory analysis of the data set on Section 4.3.1. Furthermore, some of the features used to predict purchase likelihood were directly imported from the set of features used to predict customer engagement. Nevertheless, due to the peculiarities related to customers' basket additions, we introduce new session attributes that increase prediction rates. This new angle of information is even more relevant considering the CLM data set's limitations. Moreover, within this scope of purchase likelihood prediction it would be even more advantageous

to access registered user data. Having privileged data such as user demographics, past purchases and product details could only improve the predictors' accuracy.

5.3.2 Data Set Restriction

The problem of predicting purchase likelihood is only applicable to a subset of the original CLM data set. In fact, only sessions that have engaged with purchasing, i.e. added the first item to the cart, are considered. Therefore, the tabular data that is fed to the learning models only contains rows that match requests of sessions that already have engaged purchasing, i.e. AIB user sessions.

Since the learning models will forecast purchase likelihood, predictions will only start from the moment the engagement was achieved. Therefore, from the original CLM data set, only 45 801 user sessions are considered. Given all in-session restrictions, the considerable data set amasses to a total of 3 831 712 requests.

From this pool of requests, 20% were set aside to constitute an untempered test set, used to evaluate the learning models and draw conclusions. The remaining 80% were used during the exploratory analysis phase and training of predictors. Unlike the data set used to model purchasing engagement, the AIB user session's are heavily imbalanced. Moreover, only 15% of all AIB user session's requests correspond to basket additions. For this reason, we consider two different sampling strategies to build the training set:

- **Stratified:** This subdivision is stratified, such that the percentage of requests corresponding to basket additions and all the others are approximately the same as that in the initial data set.
- **Balanced:** This subdivision samples entries from the original data set, such that the training set becomes balanced. In other words, there is approximately the same amount of requests that lead to a basket addition and requests that do not.

5.3.3 Explanatory Analysis

This section's purpose is to explore the CLM data set and present its most relevant features for purchase likelihood prediction. Furthermore, this analysis is fundamental because it allows for a deeper knowledge of the data set and its domain. During this section we will introduce every feature that takes part in the learning models. We only focus on the set of features that are used by the learning models, however, the actual explanatory analysis was more extensive as the process of feature selection requires a lot of trial and error.

Table 5.1 depicts the set of features that feed the studied techniques. These features were chosen after an extensive process of testing different hypothesis. For the remainder of this section we will briefly explain why these features add value to the classifier.

Table 5.1: The label, short description of the variables used in the model selection and the type of each variable.

Feature	Label	Type
logNBA	logarithm of the number of previous basket additions	Numerical (continuous)
ANRP	average number of requests between purchases	Numerical (continuous)
subCat	percentage impact of the number of requests related to <i>Sub Category Menu</i> within the last 10 requests	Numerical (continuous)
searchRes	percentage impact of the number of requests related to <i>Search Results</i> within the last 10 requests	Numerical (continuous)
prodPage	percentage impact of the number of requests related to <i>Product Page</i> within the last 10 requests	Numerical (continuous)
shipPay	percentage impact of the number of requests related to <i>Shipping & Payment</i> within the last 10 requests	Numerical (continuous)
markovLik	Markov likelihood value of the next request entailing a basket addition	Numerical (continuous)
markovDis	Markov pageview sequence discrimination value	Numerical (continuous)

5.3.3.1 Number of Basket Additions

It is our understanding that a key factor driving basket additions is the number of items they have already acquired. Furthermore, we believe that the user archetype that purchases items on such retailers follow similar patterns, namely, in terms of the distribution of the number of products they buy. Therefore, this section explores the hypothesis that the number of previous basket additions influences the probability of further additions.

Figure 5.1 depicts an histogram that specifies the empirical likelihood of a basket addition for all possible session moments. It should be noted that these probabilities were derived from actual empirical frequencies. This graph accredits plausibility to our hypothesis as, according to historical data, the probability of a basket addition, as the number of pageview requests increases, seems to follow a log-normal distribution.

The validity of applying the logarithmic operation on this feature, can be explained by the same motives that justified its use when we dealt with the number of pageview requests for the purchase engagement modelling (see Section 4.3.3.1). Moreover, Figure 5.2 displays a collection of plots that evaluate how well the log-normal distribution curve fits the data. Although distribution fitting is out of the scope of this thesis, we should address some important decisions. First, even though the underlying distribution is discrete, which, given its shape, should be fitted with a binomial distribution, we chose the log-normal curve because it yields a better overall fit and we are not concerned with individual probabilities but rather with its shape. This log-normal approximation shows good fitting in all evaluations except the *Q-Q Plot* which is affected by the discrete values of the original distribution, namely when comparing high order quantiles.

Predicting Purchasing Likelihood

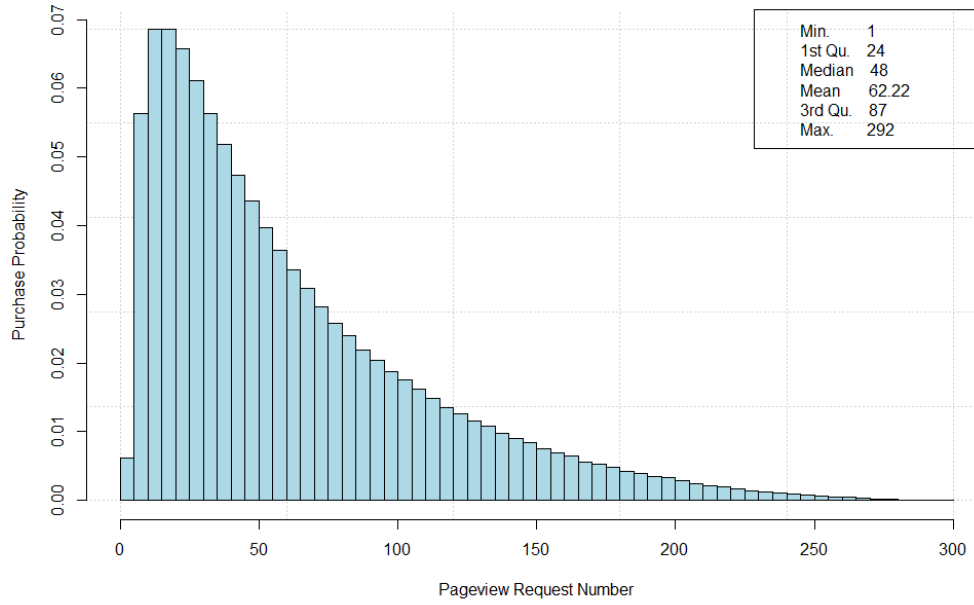


Figure 5.1: Histogram of the probability of a basket addition at different session stages (number of pageview request).

5.3.3.2 Average Number of Requests Between Purchases (ANRP)

This feature assumes that users tend to follow a purchase cadence, i.e. on average, they perform the same number of requests before adding an item to the cart. We believe that this is a valid assumption since each customer tends to follow the same search method to find products. Furthermore, e-grocery website's mimic the organization of physical stores, that is, different sections containing products from similar categories. As users try to optimize the time they spend at the store, they will keep adding products as they travel from section to section. On the website, we believe that users also try to minimize their shopping time by engaging a certain buying pattern that we want to capture.

Unlike others, this feature is specific for each session as the ANRP is updated each time the user adds an item to the basket. In order to inform the prediction models how every request stands regarding the last basket addition and the session's ANRP, we take the ratio between the number of requests since the last basket addition and the current session's ANRP. So, for instance, if the last item was added to the cart 3 requests ago, and the current ANRP is 6.7 requests, then, the value passed to the predictors would be 0.448 , which means that only 44.8% of the average number of requests were performed since the last addition.

Predicting Purchasing Likelihood

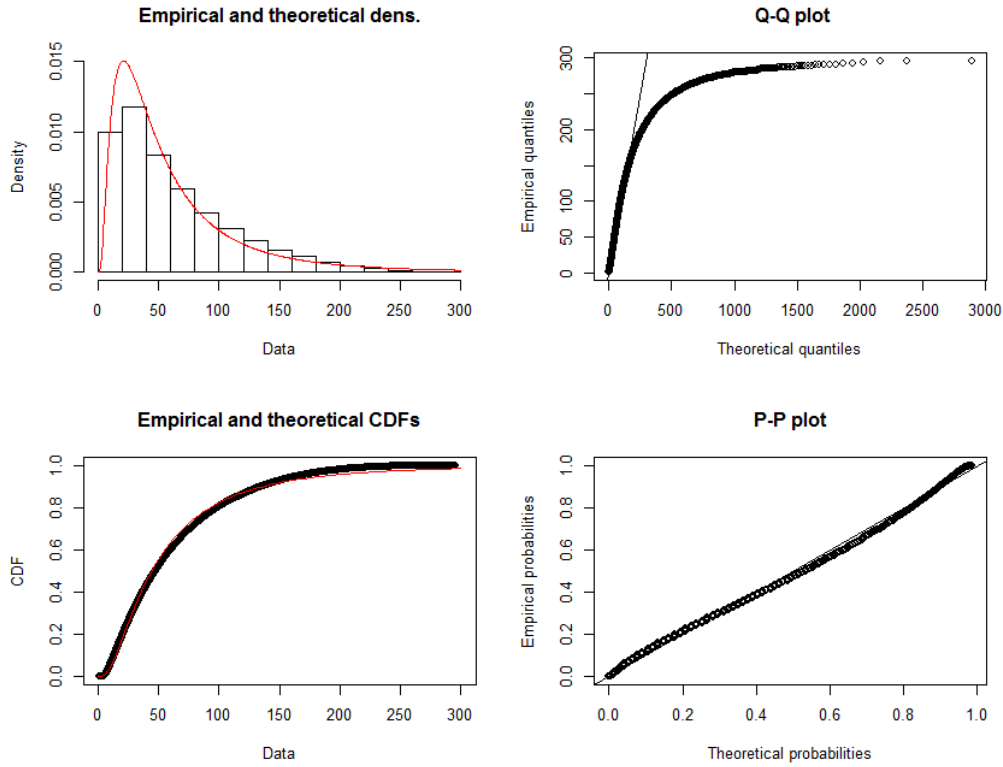


Figure 5.2: Collection of plots assessing the quality of fitting a log-normal distribution over the empirical histogram, representing the moment of purchase.

5.3.3.3 Pageview Session Span

During the AIB user sessions, that we focus on during this chapter, users tend to execute two distinct tasks: *search/purchase* and *shipping/payment*. Naturally, basket additions occur during the *search/purchase* phase thus the need to distinguish both stages. In order to do so, we resort to the concept of *pageview session prevalence* (see Section 3.3.2.2) and analyze how users' behaviour alters throughout the session. Furthermore, we use a *sliding window* strategy to evaluate the ten most recent sessions' pageview requests. During this request window, we look for specific pageview requests that pinpoint specific user behaviours:

- **Sub Category Menu:** Navigating through product categories' menus is associated with users that are still in the *search/purchase* stage.
- **Search Results:** Users who submit a search term request are also linked to the *search/purchase* behavioural pattern.
- **Product Page:** Requesting the page with detailed product information displays active interest in a certain product, thus the presence of this pageview within the 10 request sliding window displays clear *search/purchase* intentions.

- **Shipping & Payment:** During the preprocessing stage we chose to group all requests that were somehow related to the *shipping/payment* stage. Moreover, these requests relate to choosing delivery dates, reviewing basket items, handling payment, etc.

In essence, these four user behavioural indicators assume continuous values that represent the percentage of requests, within the sliding window, that match the given pageview.

5.3.3.4 Markov Purchase Likelihood

Thus far, every feature has been directly related to session metrics or specific actions. However, we can also retrieve information from the sequence of clickstreams. Therefore, we propose a measure of purchase likelihood by using first order Markov chains as specified by [MLSL04].

Markov chains, named after *Andrey Markov*, are mathematical systems that hop from one state to another based on transition probabilities. First order Markov chains, also known as *memoryless* Markov chains, are systems where the probability of the next state depends only on the current state and disregards everything that happened before.

Clickstream data is naturally adapted to be modeled by Markov chains as each session is an ordered sequence of requests. Moreover, assuming the non-randomness of these requests, Markov modelling yields detailed information regarding the probability of requesting a pageview j after pageview i .

We used AIB user session's clickstream data to build a first order Markov chain that models the transitions among different pageview requests. Since we are concerned with purchase likelihood, this feature uses the values of the transition matrix that specify the probability of the next request being a basket addition depending on the present pageview. This is a static feature as the probability of requesting a basket addition, regardless the current state, is pre-computed into the transition matrix.

5.3.3.5 Pageview Sequence Likelihood

For this feature we reuse Markov for discrimination (see Section 4.3.3.5). This technique allows for the discrimination of different Markov processes for a specific sequence of Markov states. In Section 4.3.3.5 we used this technique to associate a certain clickstream sequence with AINB or AIB user sessions. However, in the context of purchase likelihood prediction, we do not have a discrimination of different user behaviour. Nevertheless, following the line of thought behind Section 5.3.3.3, we believe that it is possible to distinguish different stages of purchasing, namely, between *search/purchase* and *shipping/payment* stages. Therefore, we split AIB user sessions so that, the requests made until the last basket addition (*search/purchase*) and the requests made after the last basket addition (*shipping/payment*) are separated. Then, we modeled Markov processes for both sets of requests, in order to obtain distinct probabilistic models that describes each set. To perform the discrimination, we used a sliding window of the last ten requests because we want to capture the change of behavior during the session. As for the discrimination, we follow the same

procedure of engagement prediction (see Section 4.3.3.5) to obtain a single factor that, for that specific window, highlights one behavioural pattern instead of the other.

5.4 Learning Models

In order to test the predictive power of selected features, we submit our data set into two different learning models: logistic regression and random forests. In this section we specify the learning process, relevant parameters and evaluation metrics.

5.4.1 Techniques and Evaluation Criteria

The prediction models that are used to estimate purchase likelihood are the same as the ones used for predicting purchasing engagement (see Section 4.4.1 and Section 4.4.2). Having no tuning parameters, the binary logistic model is simply fed with a new data set and its results are analyzed. On the other hand, random forests have two tuning parameters regarding the number of trees and the number of independent variables selected. For this study we follow the recommendations reported by [Bre01] and consider a large ensemble of trees (i.e. 1000 trees) and the truncated square root of the number of variables as the number of independent variables selected (i.e. 3 variables).

Regarding model evaluation criteria, we resort to the same four indicators that also measure the performance of purchasing engagement predictions, namely, the ROC curve and respective AUC, the precision/recall curve and the accuracy curve (see Section 4.4.3). Due to the imbalanced nature of the data set, we also evaluate each model with a measure that specifies the accuracy for each class separately. The *class accuracy rate* is a combined plot of the *true positive rate* (basket addition) and the *true negative rate* (normal request). In essence, this analysis depicts, for each possible class, the ratio between prediction values and the actual record's class label.

5.4.2 Results

Due to memory constraints, we had to limit the training data set to 50 000 entries out of the 3 831 712 available. In order to avoid overfitting, we applied a validation strategy where we sample 10 unique samples of 50 000 entries out of the whole training data set. Then, for each sample, we feed our models with 80% of the data during the training phase and test with the remaining 20%.

Figure 5.3 depicts the performance measures of both models, logistic regression and random forests, after these were trained with the imbalanced data set and tested with the reserved test set. These measurements are agnostic regarding the cut-off value that determines the frontier between entries that result in posterior basket addition and entries that do not. Moreover, the best cut-off value depends on the marketers' intentions which affect the choice of precision/recall trade-off. From the analysis of Figure 5.3 we can conclude that anonymous purchase likelihood prediction in this context is promising. The AUC values are high both when logistic regression and random forests are used. The results obtained also show that random forests outperforms

Predicting Purchasing Likelihood

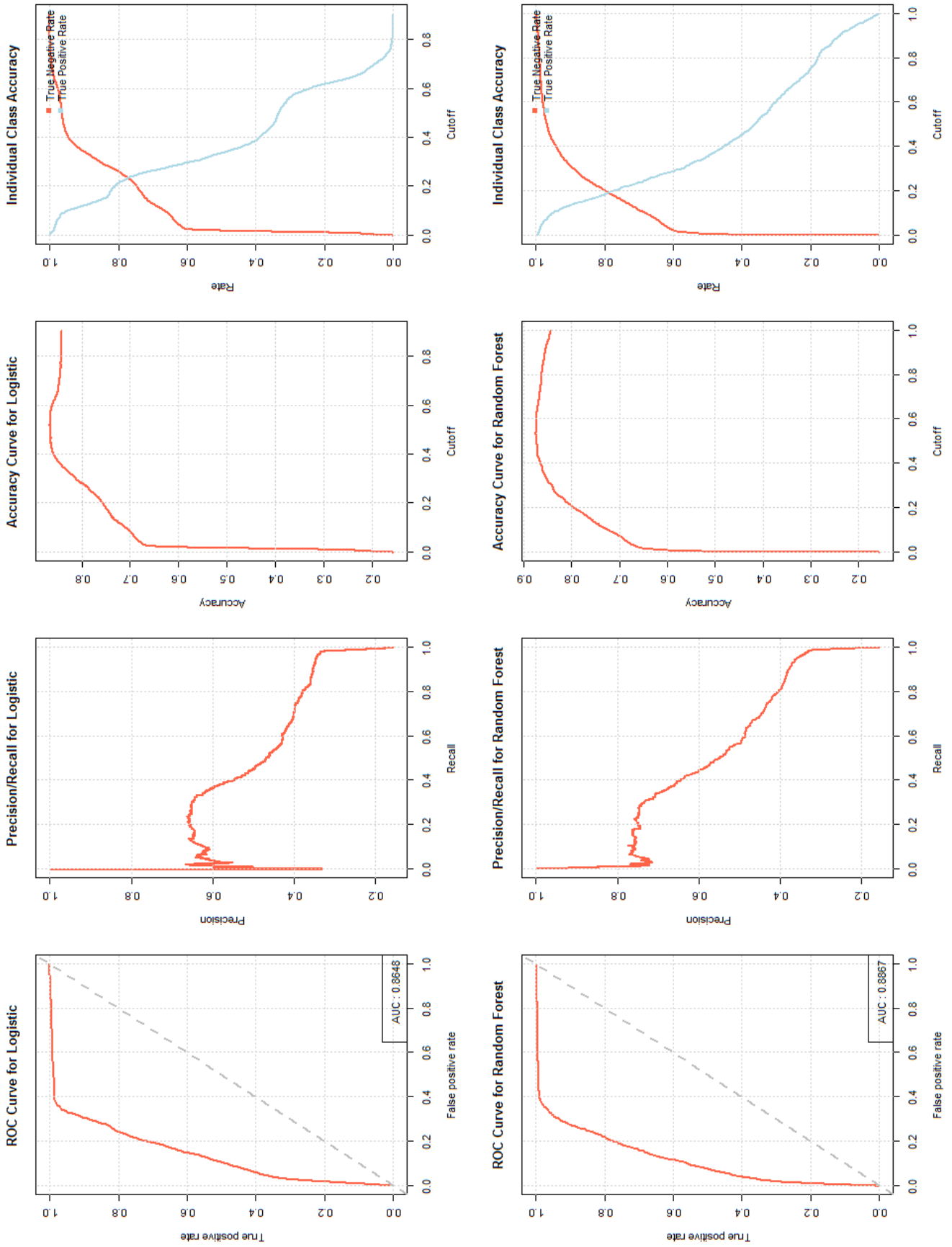


Figure 5.3: Collection of plots for different performance metrics for predicting purchase likelihood with an imbalanced data set. The top four plots refer to the evaluation of the logistic regression model, while the remaining bottom four relate to the evaluation of the random forest model.

logistic regression in all performance metrics. Moreover, the AUC value for random forests is close to 2.5% higher than the logistic regression's AUC. On the other hand, the accuracy rate for the class that corresponds to cases labeled with basket addition (positive values), is low for logistic regression and random forests. For example, a cutoff value of 0.5, yields an accuracy under 40%, for both learning models. This could be problematic if web analysts wish to target as many basket additions as possible.

Figure 5.4 depicts the performance measures of both models, logistic regression and random forests, after these were trained with the balanced data set and tested with the reserved test set. Comparing the results between balanced and imbalanced data sets, we can denote a slight decrease of overall prediction power when the models are fed with a balanced dataset. Moreover, this statement is supported by lower AUC values and lower overall accuracy levels. However, the individual class accuracy plot reveals that modelling with a balanced data set yields different prediction models. Unlike the imbalanced data set's models, we can observe that accuracy levels for the class that corresponds to cases labeled with basket addition (positive values), is much higher than before. Notwithstanding, this increase of accuracy for the positive class is balanced with a decrease of accuracy for the negative class.

Both prediction models were implemented using *R programming language* libraries.

5.5 Discussion

In this chapter we used the logistic regression and the random forests models to describe the association between general clickstream information of AIB users and whether they will add an item to the basket, given their past requests. This model provides a new tool for an e-commerce web analyst that helps to infer users' behavior, and as a result, to improve online conversion rates. Moreover, this tool displays a dynamic behavior as it follows users' sessions and continuously evaluating purchasing intentions.

The obtained results reveal that prediction models are effective in this domain. Unfortunately, the CLM data set does not contain the detailed clickstream variables and demographics that would increase the set of features and add more information to the prediction models. Moreover, we believe that any information regarding the categories and/or products that users search for could only increase prediction rates as, for instance, buying products from the *milk* section is totally different than buying products from the *light-bulbs* section. Nonetheless, the process of feature engineering revealed to be fruitful as, for instance, the random forest model is able to reach accuracy levels of 86%. Moreover, these results show that anonymous clickstream data from online grocery retailers is enough to develop a solid model for purchase engagement prediction.

From the digital marketing point of view, these results are very encouraging as new methods of targeting customers could be derived from this solution. In fact, we believe that the learning models developed within this chapter, from a business standpoint, are even more interesting than the models related with customer engagement prediction. Being able to anticipate basket additions opens new possibilities for marketing maneuvers. For example, if the model predicted a basket

Predicting Purchasing Likelihood

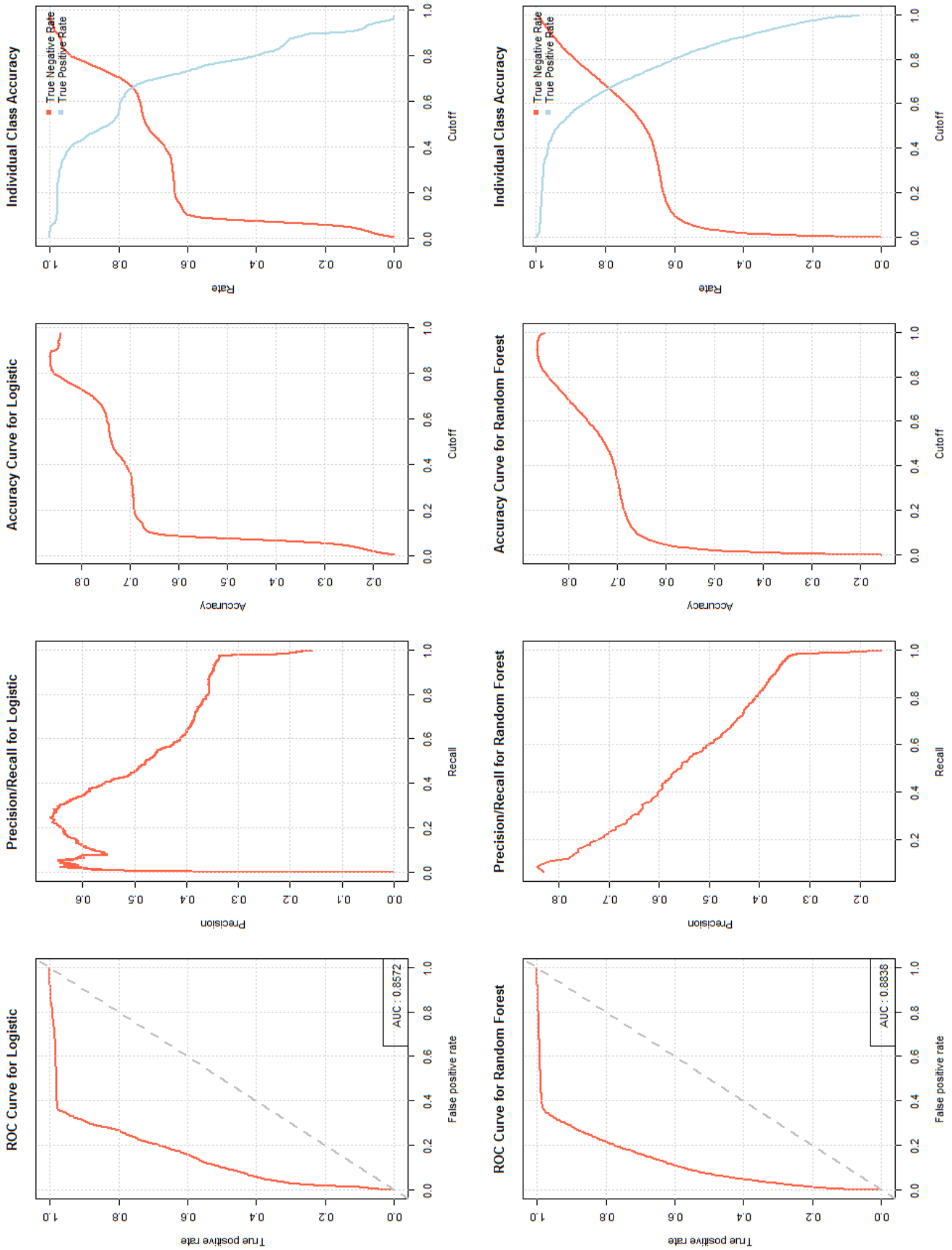


Figure 5.4: Collection of plots for different performance metrics for predicting purchase likelihood with a balanced data set. The top four plots refer to the evaluation of the logistic regression model, while the remaining bottom four relate to the evaluation of the random forest model.

addition, the web server could adjust the response to the user so that more profitable products were given more emphasis. Furthermore, the web analyst is able to deploy different training sets, balanced or imbalanced, according to the objectives of his campaign. As concluded, imbalanced data sets yield a model with better model to target requests that do not entail basket additions, while balanced data sets favor the identification of requests that anticipate purchases.

There are alternative approaches including neural networks, deep learning, categorical principal analysis and further decision tree-based methods which can be used to find the *best* approach for predicting customer engagement. Choosing the best approach, which is extensively discussed in the literature, would go beyond the scope of this thesis, and is not our aim [K⁺95, VV98]. The logistic regression was chosen here for comparison reasons as most of our focus was dedicated to the random forests model. The prediction model developed with random forests is solid, and would be suitable for an extension of features if the CLM data set was also extended to include registered user information.

Chapter 6

Conclusions

This thesis originated from practical business questions related with a major European food retailer with online presence. The main purpose was to investigate statistical approaches on clickstream data, as the aggregated sequence of pageview requests executed by a particular user, and other user navigation features, can provide insight into their intentions, specifically with respect to purchase engagement and real-time purchase likelihood prediction.

6.1 Clickstream Data Preprocessing

We worked upon one major data source that was retrieved from the retailer's server logs. The log file contains information about every single click made by a user on a web browser while surfing the Internet, corresponding to an HTTP request sent to the website's server. The log data that we used for this study, referred to as the *CLM data set*, was not complete, in the sense that some relevant clickstream attributes were missing. One limitation for the analysis of the CLM data set is the lack of knowledge regarding page references. When a user returns to a page that has already been visited (downloaded from the server) during the same session, the second access to that page will, in most cases, result in viewing the previously downloaded version of the page without sending a request to the server, due to client-side browser caching. This problem could be solved by knowledge of the website's structure. However, if the website is designed in such way that pageview requests do not have to follow a rigid sequence, i.e. one can request any pageview from every page of the website, structural website knowledge is not enough. Indeed, the website behind the CLM data set displays this dynamic behavior thus making it impossible to track browser caching events. Throughout this study we assumed that sessions could not cache previous requests. There is no guarantee that this assumption is true but we believe that it does not impact the learning models.

The original CLM data set, as well as general web log files, do not consist of well structured data and cannot be used directly for analytical purposes. Making clean clickstream data that can

Conclusions

provide reliable information about web browsing behavior requires a good understanding of the domain. For this reason we dedicate a relevant part of this research to explain every stage and technique related to *data preprocessing* (see Section 2.4.2). For example, we highlight the need to remove redundant records of log file data when a user requests a web page containing media files, as the request results in several records/lines in the web log file that represent just one page request. On the other hand, we also removed records in the web log files made by bots, as those lines do not reflect human navigational behavior.

For analytical purposes, we excluded sessions that performed five or less pageview requests, also known as *bounce visits*. These short sessions were not taken into account because, given the main purpose of studying purchasing intentions, the website's structure does not allow purchases with such few requests. A major part of data preprocessing is related with *sessionization*, i.e. determining when the user starts and ends a visit to the website. This procedure is not exact because the user can request one or more pageviews and leave the browser open. If the user then returns to the website through the open page in the browser, the sessionization strategy should decide whether to consider that activity a continuation of the previous session or a new session. We used a web sessionization rule of splitting the session into two when the time between two clicks is more than 10 minutes. Regarding *user identification*, another preprocessing stage, we assume that each session is performed by a new user even though the user might have originated previous sessions. This decision relates to the nature of the website from which the CLM data set is from, that is, an online grocery retailer, where, we believe, people tend to purchase items on a monthly basis, the exact period comprehended in the CLM data set. Another, very important, preprocessing stage relates to the identification of pageviews, i.e. label each possible request with a tangible action. During pageview identification we had to use specific company knowledge, nevertheless, we believe that our models remain agnostic, specially because we tagged pageviews as generally as possible.

Arguably the major drawback of the CLM data set, is the void of information regarding registered user data. This type of data relates to users that are registered in the company's databases, that store privileged user information such as demographics and historical purchases. Having access to this sensitive information would allow for the development of more sophisticated features capable of a better characterization of the data set. On the other hand, the lack of information regarding registered user data makes our work more prone to generalization, namely considering other e-groceries, because we are working with as little company knowledge as possible.

6.2 Clickstream Data Exploratory Analysis

Throughout this thesis we have shown how clickstream data is able to provide an insight into visitor's behavior. In order to craft relevant features we reviewed some important web metrics and statistical reporting using clickstream data. Depending on the goals of the analysis, this data could be transformed and aggregated or segmented at different levels of abstraction to provide useful

metrics. For instance, these metrics could be reported at the level of website or user session. In Chapter 3 we illustrated some of these metrics and how valuable is this analysis.

One of the most important analysis we performed, regards the pageview segmentation of user sessions with the introduction of *pageview session prevalence* and *pageview session span* reports. Both these reports delve into user sessions and reveal how different user archetypes spread the total amount of requests into each identified pageview. Using these methods, we were able to highlight specific pageviews that are able to split different user archetypes. For example, we discovered that ANIB users are four times more likely to perform actions related to their personal profiles than AIB users. Moreover, this particular distinction was identified by comparing ANIB and AIB pageview session spans.

We also performed an exploratory analysis on depth-of-visit metrics, by the number of pages requested and the number of products added to the basket. These metrics revealed to be very important because they are the main source of information for the learning models to perceive the stage of the session.

Additionally, we explored other features that provided new information regarding the sequence of pageviews. Unlike other features that focus on session metrics or specific actions, we also extracted information from the sequence of requests that each session performs. Therefore, we applied a measure of pageview sequence similarity by using Markov for discrimination. In essence, this technique allows the quantification of the degree of similarity between a particular session and two behavioural patterns. We apply this concept, for instance, when we want to know whether a certain sequence of requests is more likely to belong to an AIB or ANIB user.

Overall, we need to stress that exploratory analysis was a fundamental part of this research as we dedicated most of our time into understanding, segmenting and plotting the CLM data set. If new sources of data were to become available, we would go back to this exploratory phase and mine the most relevant and descriptive features using the same methods.

6.3 Predicting Purchasing Engagement

One of the main objectives of this thesis was to create a binary classification model capable of predicting visitor purchasing engagement. We used the logistic regression and the random forests models to describe the association between general clickstream information concerning visits and whether a visitor will engage in online purchasing behavior during his visit to the website. In practice, we concluded that purchasing engagement prediction is the same as predicting whether visitors will become AIB or ANIB users.

These models were fed with features mined from the exploratory analysis. Both learning models allows us to obtain conditional probability estimates of a first basket addition. This way, the web owner will be able to identify high potential visitors in terms of conversion tendency, and generate leads for suitable targeting actions. In a web-focused marketing solution, the targeting action might help retaining the customer and avoid defection to competitors.

Conclusions

In order to assess each classification model prediction capabilities, we used three evaluation methods that are agnostic in regards to the cutoff value, i.e. the number that defines the border between AIB and ANIB users: the ROC curve (with its respective AUC value), a precision/recall curve and the model's accuracy curve. The results show that, for the e-grocery domain, prediction models are very effective. In fact, the best model, random forests, was able to output an AUC value of *0.8415* and accuracy levels above *70%* for most of the possible cutoff values.

The models can be improved by considering more predictors, for example, any knowledge about the web pages viewed during the session. As previously stated (see Section 6.2), using registered users who login to the website, depending on the requested information in the registration forms, may increase the predictive performance of the models. This includes demographical (gender, age, occupation, etc.) and historical (past purchases, average spending, average session time, etc) attributes.

6.4 Predicting Purchase Likelihood

The other main objective of this thesis, arguably the most relevant, was to create a binary classification model capable of continuously predicting when AIB users are going to add an item to their virtual basket. We used the logistic regression and the random forests models to describe the association between general clickstream information concerning AIB users' sessions and whether a they will perform any basket addition.

From a business standpoint, we believe these models are extremely relevant because they predict events that are directly related to revenue. Moreover, the web owner is able to anticipate product basket additions and act according to users' intentions. In order to stimulate customer profitability the web owner could deploy measures that, for example, would highlight products with higher margins, in case the models reveal a high probability of basket addition.

Unlike the data set that fed the model of purchasing engagement prediction, the data set containing all the requests from AIB users (the segment that is mine-able for purchase likelihood prediction) is imbalanced. In fact, only *15%* of the requests correspond to basket additions. With this in mind, we developed two sampling strategies to produce different training sets for our models: one that sampled records and kept the original class balance, and another where we balanced the presence of both classes. These strategies yielded distinct results. Training with a sample of the original data set yielded models with better AUC (*0.8867* with random forests) and overall accuracy levels (above *80%* with random forests) but, compared to the models trained with a balanced data set, they have a much lower individual positive class accuracy (adding an item to the basket). Determining which is the best model depends on the web owner's objectives. For example, if the objective is to target as most basket additions as possible then, the best model is the one trained with a balanced data set. If, on the other hand, the objective is to target basket additions with as few errors as possible then, the model trained with the original data set is better suited.

Similarly to the models for predicting purchasing engagement, these binary classifiers could be improved if more predictors, from additional sources of information, were introduced.

6.5 Future Research

As we have been stating throughout this thesis, obtaining valuable data is the most important aspect in order to build a solid prediction model. Moreover, regarding online retailing clickstream prediction, we believe that a complete data set, with privileged attributes from registered users, opens a whole new dimension of insightful features that could improve prediction quality. We believe that, in a near future, companies ought to explore this line of thought if they want to remain competitive.

On the other hand, future work could be developed in applying new learning algorithms to fit clickstream data, namely, by introducing other models such as neural networks, support vector machines, genetic algorithms, etc. Along side these models, further evaluation criteria and even model selection methods could be studied in order to better understand which options is more appropriate for each occasion.

Finally, we wish to stress that, given that the field of clickstream data research is still in its infancy, much research still needs to be done. With the insurgence of new and faster technology, the concept of *big data* is very hot at the moment, specially because companies can, like never before, translate customer data into higher revenue. This study was the first, to our knowledge, to investigate online customer's intentions within the online food retailing domain. We believe that future literature will approach this topic again and improve our results by introducing new methodologies for preprocessing, feature engineering and modelling.

Conclusions

References

- [AB02] Rizal Ahmad and Francis Buttle. Customer retention management: A reflection of theory and practice. *Marketing Intelligence & Planning*, 20(3):149–161, 2002.
- [Agr96] Alan Agresti. *An introduction to categorical data analysis*, volume 135. Wiley New York, 1996.
- [Ber83] Leonard L Berry. Relationship marketing. American Marketing Association, 1983.
- [BGL⁺09] M Boucadair, JL Grimault, P Lévis, A Villefranque, and P Morand. Anticipate ipv4 address exhaustion. *A Critical Challenge for Internet Survival*, 2009.
- [BHS06] Erik Brynjolfsson, Yu Jeffrey Hu, and Michael D Smith. From niches to riches: Anatomy of the long tail. *Sloan Management Review*, 47(4):67–71, 2006.
- [BLA⁺02] Randolph E Bucklin, James M Lattin, Asim Ansari, Sunil Gupta, David Bell, Eloise Coupey, John DC Little, Carl Mela, Alan Montgomery, and Joel Steckel. Choice and the internet: From clickstream to research stream. *Marketing Letters*, 13(3):245–258, 2002.
- [BMP11] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.
- [BRD⁺00] Sebastiano Biondo, Emilio Ramos, Manuel Deiros, Juan Martí Ragué, Javier De Oca, Pablo Moreno, Leandre Farran, and Eduardo Jaurieta. Prognostic factors for mortality in left colonic peritonitis: a new scoring system. *Journal of the American College of Surgeons*, 191(6):635–642, 2000.
- [Bre96] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
- [Bre01] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [Bro14] Jason Brownlee. Neglected machine learning ideas. <https://scottlocklin.wordpress.com/2014/07/22/neglected-machine-learning-ideas/>, 2014. (Visited on 06/08/2015).
- [BS09] R. E. Bucklin and C. Sismeiro. Click here for internet insight: Advances in click-stream data analysis in marketing. *Journal of Interactive Marketing*, 23(1):35–48, 2009.
- [BT04] Philip Bligh and Douglas Turk. *CRM unplugged: releasing CRM’s strategic value*. John Wiley & Sons, 2004.

REFERENCES

- [Cas05] Carlos Castillo. Effective web crawling. In *ACM SIGIR Forum*, volume 39, pages 55–56. ACM, 2005.
- [CD08] Dumitru Ciobanu and Claudia Elena Dinucă. A new method for session identification in clickstream analysis. *Recent Researches in Tourism and Economic Development ISBN*, pages 978–1, 2008.
- [CK04] Yoon Ho Cho and Jae Kyeong Kim. Application of web usage mining and product taxonomy to collaborative recommendations in e-commerce. *Expert systems with Applications*, 26(2):233–246, 2004.
- [CKLT03] Kwok-Wai Cheung, James T Kwok, Martin H Law, and Kwok-Ching Tsui. Mining customer product ratings for personalized marketing. *Decision Support Systems*, 35(2):231–243, 2003.
- [CKT97] Dennis R Capozza, Dick Kazarian, and Thomas A Thomson. Mortgage default in local markets. *Real Estate Economics*, 25(4):631–655, 1997.
- [CLA⁺03] Dan Cosley, Shyong K Lam, Istvan Albert, Joseph A Konstan, and John Riedl. Is seeing believing?: how recommender system interfaces affect users’ opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 585–592. ACM, 2003.
- [Cli12] Brian Clifton. *Advanced web metrics with Google Analytics*. John Wiley & Sons, 2012.
- [CMS97] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Web mining: Information and pattern discovery on the world wide web. In *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on*, pages 558–567. IEEE, 1997.
- [CMS99] R Cooley, Bamshad Mobasher, and J Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1(1):5–32, 1999.
- [Cox58] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 215–242, 1958.
- [CP03] Injazz J. Chen and Karen Popovich. Understanding customer relationship management (crm). *Business Process Management Journal*, 9(5):672–688, 2003.
- [CTH⁺10] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, Series Bridging the gap between physical location and online social networks, pages 119–128, Copenhagen, Denmark, 2010. ACM.
- [CTS00] Robert Cooley, Pang-Ning Tan, and Jaideep Srivastava. Discovery of interesting usage patterns from web data. In *Web Usage Analysis and User Profiling*, pages 163–182. Springer, 2000.
- [DC11] CE Dinuca and D Ciobanu. Improving the session identification using the mean time. 2011.

REFERENCES

- [DDL11] Thomas H. Davenport, Leandro DalleMule, and John Lucker. Know what your customers want before they do. *Harvard Business Review*, pages 84–92, 2011.
- [DEKM98] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.
- [DNB02] Devanshu Dhyani, Wee Keong Ng, and Sourav S Bhowmick. A survey of web metrics. *ACM Computing Surveys (CSUR)*, 34(4):469–503, 2002.
- [Dom12] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.
- [Dou92] Comer Douglas. Internetworking with tcp/ip: Principles, protocols, and architecture. *fourth edition*, 1, 1992.
- [EY10] NM Abo El-Yazeed. An overview of preprocessing of web log files for web usage mining. *Online] Available: etms eg. org/cms/upload/1354650330. pdf*, pages 1–16, 2010.
- [FGM⁺99] Roy Fielding, Jim Gettys, Jeffrey Mogul, Henrik Frystyk, Larry Masinter, Paul Leach, and Tim Berners-Lee. Hypertext transfer protocol–http/1.1, 1999.
- [FPSS96] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.
- [HB12] Antonio Hyder and Enrique Bigné. Does web site engagement lead to making a purchase? 2012.
- [HKP06] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining, southeast asia edition: Concepts and techniques*. Morgan kaufmann, 2006.
- [HLC⁺84] Frank E Harrell, Kerry L Lee, Robert M Califf, David B Pryor, and Robert A Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 3(2):143–152, 1984.
- [HM82] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.
- [Ho95] Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.
- [Jam11] Mohammadamin Jamalzadeh. *Analysis of clickstream data*. PhD thesis, Durham University, 2011.
- [JZFF10] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Recommender systems: an introduction*. Cambridge University Press, 2010.
- [K⁺95] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145, 1995.
- [Kan11] Mehmed Kantardzic. *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

REFERENCES

- [Kau07] Avinash Kaushik. *Web Analytics: An Hour A Day (W/Cd)*. John Wiley & Sons, 2007.
- [KJSH06] Su-Yeon Kim, Tae-Soo Jung, Eui-Ho Suh, and Hyun-Seok Hwang. Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert systems with applications*, 31(1):101–107, 2006.
- [KMPZ04] Ron Kohavi, Llew Mason, Rajesh Parekh, and Zijian Zheng. Lessons and challenges from mining retail e-commerce data. *Machine Learning*, 57(1-2):83–113, 2004.
- [KMS04] Alexander H Kracklauer, D Quinn Mills, and Dirk Seifert. Customer management as the origin of collaborative customer relationship management. In *Collaborative Customer Relationship Management*, pages 3–6. Springer, 2004.
- [KND13] Ankit R Kharwar, Ni A Naik, and Niyanta K Desai. A complete pre processing method for web usage mining. 2013.
- [LC04] Shu-Hsien Liao and Yin-Ju Chen. Mining customer knowledge for electronic catalog marketing. *Expert Systems with Applications*, 27(4):521–532, 2004.
- [Les15] Jure Leskovec. New directions in recommender systems. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 3–4. ACM, 2015.
- [Liu07] Bing Liu. *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media, 2007.
- [LSY03] Greg Linden, Brent Smith, and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, 2003.
- [LT07] Kenneth C Laudon and Carol Guercio Traver. *E-commerce*. Pearson/Addison Wesley, 2007.
- [LY04] Feng Li and Irene Yousept. Online supermarkets: emerging strategies and business models in the uk. *BLED 2004 Proceedings*, page 30, 2004.
- [LYL96] Hongjun Lu, Sam Yuan, and Sung Ying Lu. On preprocessing data for effective classification. In *ACM SIGMOD’96 Workshop on Research Issues on Data Mining and Knowledge Discovery*. Citeseer, 1996.
- [McE02] Noelle McElhatton. Case study - tesco. *Direct Response*, pages 33–34, 2002.
- [MF04] Wendy W Moe and Peter S Fader. Dynamic conversion behavior at e-commerce sites. *Management Science*, 50(3):326–335, 2004.
- [MLSL04] Alan L Montgomery, Shibo Li, Kannan Srinivasan, and John C Liechty. Modeling online browsing and path analysis using clickstream data. *Marketing Science*, 23(4):579–595, 2004.
- [MVdPCeC12] Vera L Miguéis, Dirk Van den Poel, Ana S Camanho, and João Falcão e Cunha. Predicting partial customer churn using markov for discrimination for modeling first purchase sequences. *Advances in Data Analysis and Classification*, 6(4):337–353, 2012.

REFERENCES

- [Nga05] EWT Ngai. Customer relationship management research (1992-2002) an academic literature review and classification. *Marketing Intelligence & Planning*, 23(6):582–605, 2005.
- [NWW14] Bert Nagelvoort, Aad Weening, and Richard van Welie. European b2c e-commerce report 2014. Technical report, Ecommerce Europe, 2014.
- [NXC09] E. W. T. Ngai, L. Xiu, and D. C. K. Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2):2592–2602, 2009.
- [Pal02] Jonathan W Palmer. Web site usability, design, and performance metrics. *Information systems research*, 13(2):151–167, 2002.
- [PC09] Jungkun Park and Hoeun Chung. Consumers’ travel website transferring behaviour: analysis using clickstream data-time, frequency, and spending. *The Service Industries Journal*, 29(10):1451–1463, 2009.
- [Pen13] Joana Margarida Caldas da Silva Penim. Online grocery shopping: an exploratory study of consumer decision making processes. 2013.
- [Pin99] B Joseph Pine. *Mass customization: the new frontier in business competition*. Harvard Business Press, 1999.
- [PL08] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.
- [PP10] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 1320–1326, 2010.
- [RDTM79] John S Ramberg, Edward J Dudewicz, Pandu R Tadikamalla, and Edward F Mykytka. A probability distribution and its uses in fitting data. *Technometrics*, 21(2):201–214, 1979.
- [Ret07] Industry Retailer. Average session times grow for most big retailers - internet retailer. <https://www.internetretailer.com/2007/09/05/average-session-times-grow-for-most-big-retailers>, 2007. (Visited on 06/01/2015).
- [RF03] Jr. Romano, NicholasC and Jerry Fjermestad. Electronic commerce customer relationship management: A research agenda. *Information Technology and Management*, 4(2-3):233–258, 2003.
- [RJR12] Sheetal A Raiyani, Shailendra Jain, and Ashwin G Raiyani. Advanced preprocessing using distinct user identification in web log usage data, 2012.
- [RS00] Frederick F Reichheld and Phil Schefter. E-loyalty. *Harvard business review*, 78(4):105–113, 2000.
- [RSK13] C Ramya, KS Shreedhara, and G Kavitha. Preprocessing: A prerequisite for discovering patterns inweb usage mining process. *International Journal of Information and Electronics Engineering*, 3(2):196, 2013.

REFERENCES

- [SKR01] J Ben Schafer, Joseph A Konstan, and John Riedl. E-commerce recommendation applications. In *Applications of Data Mining to Electronic Commerce*, pages 115–153. Springer, 2001.
- [SS08] Judy E Scott and Carlton H Scott. Online grocery order fulfillment tradeoffs. In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, pages 90–90. IEEE, 2008.
- [SS13] A. Surya and D. K. Sharma. An approach for web page ordering using user session. *2013 Ieee Conference on Information and Communication Technologies (Ict 2013)*, pages 1009–1013, 2013.
- [Syn14] SyndicatePlus. The online grocery shopper. Technical report, SyndicatePlus, September 2014 2014.
- [TSK⁺06] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, et al. *Introduction to data mining*, volume 1. Pearson Addison Wesley Boston, 2006.
- [TSSW10] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welp. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.
- [VdPB05] Dirk Van den Poel and Wouter Buckinx. Predicting online-purchasing behaviour. *European Journal of Operational Research*, 166(2):557–575, 2005.
- [Ver12] Ruud Verheijden. Predicting purchasing behavior throughout the clickstream. Master’s thesis, Technical University of Eindhoven, 2012.
- [VMKR13] Chintan R Varnagar, Nirali N Madhak, Trupti M Kodinariya, and Jayesh N Rathod. Web usage mining: A review on process, methods and techniques. In *Information Communication and Embedded Systems (ICICES), 2013 International Conference on*, pages 40–46. IEEE, 2013.
- [VV98] Vladimir Naumovich Vapnik and Vlamimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.